# A STOCHASTIC MARKOV CHAIN MODEL TO DESCRIBE CANCER METASTASIS

by

Jeremy Mason

---

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements of the Degree
DOCTOR OF PHILOSOPHY
(MECHANICAL ENGINEERING)

18 December 2013

# Dedication

*To my grandparents.*

*Good. Better. Best.*

# Acknowledgements

I would like to thank my advisor, Professor Paul Newton, for his continued support throughout this process and for his great and joyous personality. Thank you for creating an environment where we can both be ourselves and learn and grow at the same time. I could not have chosen a better advisor.

I would also like to thank Professors Eva Kanso, Paul Macklin, Roger Ghanem, and Timothy Pinkston for their support and suggestions regarding my research and for taking the time to be a part of my dissertation committee.

Thank you to the Bill and Melinda Gates Foundation for providing me with a means of extending my academic career to an elite level at two of the finest institutions in the nation.

Another thank you to the great people I met at USC and through USC over the past years; the AME department, the GAPP office, CED, my lab mates, my friends, and my colleagues. For being my basketball buddies and my food testers. To my family spread out across the country, thank you for being my support through it all and continually reminding me why, and that I am not alone.

And finally, to Quentin and Candace. A very special thank you for being there for me at every turning moment, for being my solid ground when things got tough, for showing me who I truly am, and for letting me show it to the world through my work and my hobbies. I could not have done it without either of you.

# Table of Contents

# List of Figures

# List of Tables

# Abstract

A stochastic Markov chain model for metastatic progression is developed for primary 8 major cancer types based on a network construction of metastatic sites with dynamics modeled as an ensemble of random walkers on the network. We calculate a transition matrix for each primary cancer and use it to construct a circular bi-directional network of primary and metastatic locations based on postmortem tissue analysis of 3827 autopsies on untreated patients documenting all primary tumor locations and metastatic sites from this population [19]. The resulting 50 potential metastatic sites are connected by directed edges with distributed weightings, where the site connections and weightings are obtained by calculating the entries of an ensemble of transition matrices so that the steady-state distribution obtained from the long-time limit of the Markov chain dynamical system corresponds to the ensemble metastatic distribution obtained from the autopsy dataset. We condition our search for a transition matrix on an initial distribution of metastatic tumors obtained from the dataset. Through an iterative numerical search procedure, we adjust the entries of a sequence of approximations until a transition matrix with the correct steady-state is found (up to a numerical threshold). Once the transition matrix for a given cancer type is computed, our metastatic progression model is based Monte Carlo simulations of collections of random walkers all leaving the primary tumor location and executing a random walk across the directed graph from site to site. The model allows us to simulate and quantify disease progression pathways and timescales of progression from the primary tumor location to other sites. Pathway diagrams are created that classify metastatic tumors as 'spreaders'

or ‘sponges’ and quantifies three types of multidirectional mechanisms of progression: (i) self-seeding of the primary tumor, (ii) reseeding of the primary tumor from a metastatic site (primary reseeding), and (iii) reseeding of metastatic tumors (metastasis reseeding). The entire process is replicated for additional primary tumors in the dataset of [19] for individual analysis and comparative purposes.

A second contribution of this work is to introduce a quantitative notion of ‘metastatic entropy’ for cancer and use it to compare the complexity and predictability associated with the 12 most common cancer types worldwide. We apply these notions of entropy and predictability directly to the autopsy dataset used to create our Markov model. The raw data, which contains the number of metastases found at all of the anatomical sites for each cadaver (whose primary tumor location is also recorded), is normalized (yielding their empirical distribution) so that we can interpret the histograms as probability mass functions (PMFs) representing the large scale (whole body) metastatic ‘signature’ of each primary cancer. We characterize the power-law distributions associated with metastatic tumor distributions for each primary cancer type. Then we calculate the entropy associated with each and use the Kullback-Liebler divergence (relative-entropy) to compare each cancer type with all of the data aggregated into an ‘all cancer’ category, whose entropy value is used as a benchmark for comparisons.

Chapter 1

# Introduction

## 1.1 Background

The identification of circulating tumor cells (CTCs) in the human circulatory system dates back to Ashworth's 1869 paper [3] in which he identified and pointed out the potential significance of cells similar to those found in the primary tumor of a deceased cancer patient. Since then, there has been sporadic focus on CTCs as a key diagnostic tool in the fight against cancer, based mostly on the so-called 'seed-and-soil' hypothesis [23, 62, 77] of cancer metastasis, in which the CTCs play the role of seeds which detach from the primary tumor, disperse through the bloodstream, and get trapped at various distant sites (typically small blood vessels of organ tissues), then, if conditions are favorable, extravasate, form metastases, and subsequently colonize. The metastatic sites offer the soil for potential subsequent growth of secondary tumors. Paget's 1889 seed-and-soil hypothesis [62] asserts that the development of secondary tumors is not due to chance alone, but depends on detailed interactions, or cross-talk, between select cancer cells and specific organ micro-environments. In 1929, J.

Ewing challenged the seed-and-soil hypothesis [22] by proposing that metastatic dissemination occurs based on purely mechanical factors resulting from the anatomical structure of the vascular system, a proposal that is now known to be too simplistic an explanation for the metastatic patterns that are produced over large populations. While the seed-and-soil hypothesis remains a bedrock theory in cancer research, it has been significantly refined over the years to incorporate our current level of understanding on how the ability for a tumor cell to metastasize depends on its complex interactions with the homeostatic factors that promote tumor cell growth, cell survival, angiogenisis, invasion, and metastasis [23].

A schematic diagram associated with the metastatic process is shown in Figure 1.1. Here, the primary tumor (from which the CTCs detach) is located in the lower part of the diagram and the distant potential secondary locations where CTCs get trapped and form metastases are shown. In this study, we will not be concerned with extravasation, colonization and the formation of secondary tumors which are complex processes in their own right [77], but rather with a probabilistic description of metastatic progression from primary neoplasm to metastatic sites; hence, we provide a quantitative framework for charting the time-evolution of cancer progression along with a stochastic description of the complex interactions of these cells with the organ micro-environments. Also shown in the figure are representative scales of a typical red blood cell ($8\mu m$), capillary diameter ($5 - 8\mu m$), CTC ($20\mu m$), and human hair diameter ($100\mu m$). The total number of remote sites at which metastases are found for any given type of primary cancer is relatively small (see the autopsy dataset described in [19]), say on the order of 50 locations, those sites presumably being the locations at which

CTCs get trapped and subsequently colonize. For any individual making up the ensemble, of course, the number of sites with metastatic tumors would be much smaller. A 'ballpark' estimate, based on the ratio of mets to primaries (from [19]) suggests a number around $9484/3827 \sim 2.5$, although in the modern era, this number is probably higher. A reasonably thorough overview of this process is described in [70].

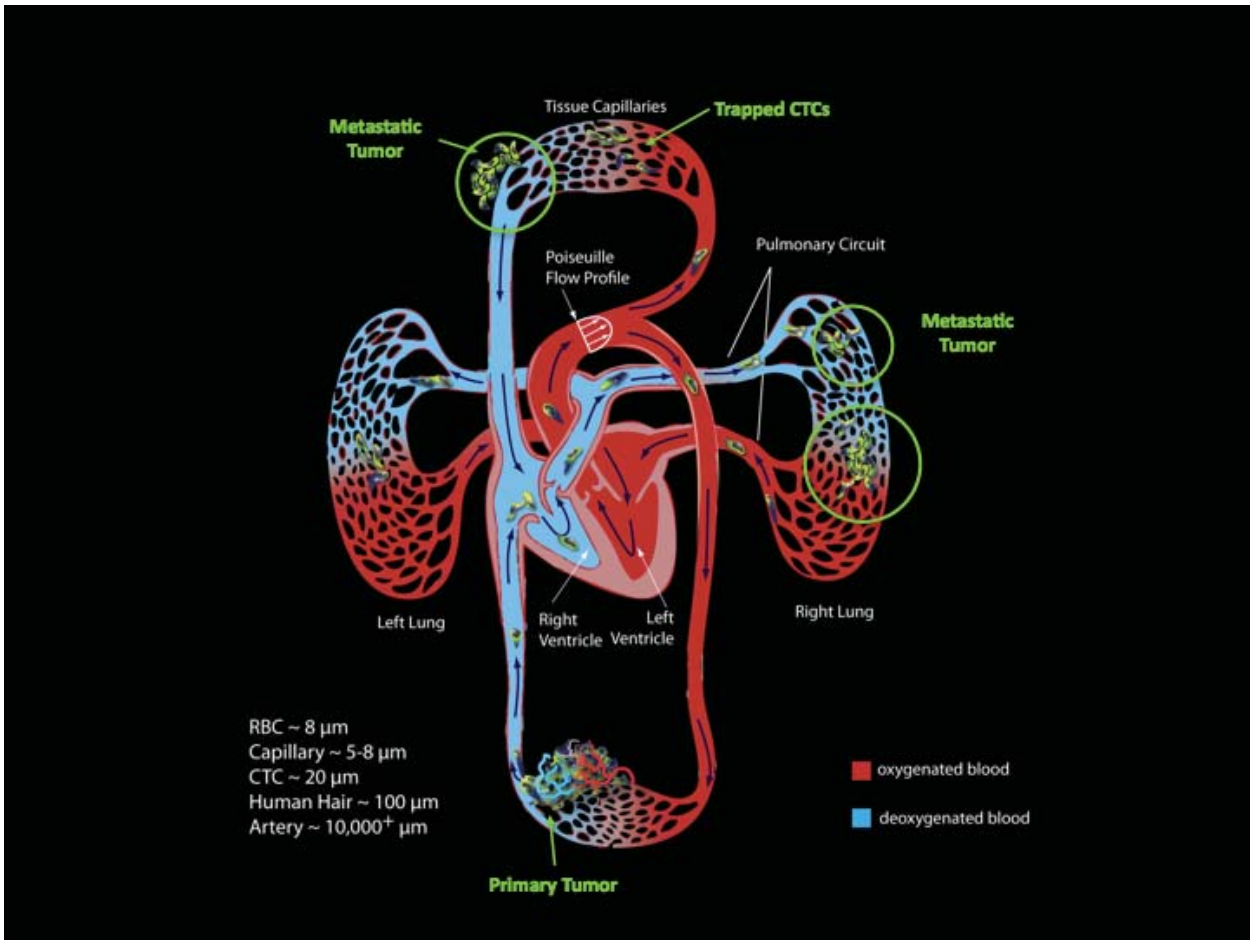


Figure 1.1: Schematic diagram of human circulatory system
Schematic diagram of human circulatory system showing circulating tumor cells (CTCs) detaching from primary tumor and getting trapped in capillary beds and other potential future metastatic locations as outlined by the 'seed-and-soil' framework.

It wasn't until recently, however, that important technological developments in the ability to identify, isolate, extract, and genetically and mechanically study CTCs from cancer patients became available (see, for example [14, 15, 34, 45, 46, 61, 63, 72]). These new approaches, in turn, produced the need to develop quantitative models which can predict/track CTC dispersal and transport in the circulatory and lymphatic systems of cancer patients for potential diagnostic purposes. As a rough estimate, data (based primarily on animal studies) shows that within 24 hours after release from the primary tumor, less than 0.1% of CTCs are still viable, and fewer than those, perhaps only a few from the primary tumor, can give rise to a metastasis. There are, however, potentially hundreds of thousands, millions, or billions of these cells detaching from the primary tumor continually over time [8, 79], and we currently do not know how to deterministically predict which of these cells are the future seeds, or where they will take root. All of these estimates, along with our current lack of detailed understanding of the full spectrum of the biological heterogeneity of cancer cells, point to the utility of a statistical or probabilistic framework for charting the progression of cancer metastasis. This is a particularly important step for any potential future comprehensive computer simulation of cancer progression, something not currently feasible. Although the dispersion of CTCs is the underlying dynamical mechanism by which the disease spreads, the probabilistic framework obviates the need to model all of the biomechanical features of the complex processes by which cells journey through the vascular/lymphatic system. The mathematical/computational framework for such an approach is provided later in detail.

## 1.2 The Markov model

We develop a new Markov chain based model of metastatic progression for primary lung cancer and for 7 other cancer types (breast, cervical, colorectal, pancreatic, prostate, ovarian, and skin), which offers a probabilistic description of the time-history of the disease as it unfolds through the metastatic cascade [77]. The Markov chain is a dynamical system whose state vector is made up of all potential metastatic locations identified in the dataset described in [19] (defined in Table 1.1), with normalized entries that can be interpreted as the time-evolving (measured in discrete steps k) probability of a metastasis developing at each of the sites in the network. One of the strengths of such a statistical approach is that we need not offer specific biomechanical, genetic, or biochemical reasons for the spread from one site to another, those reasons presumably will become available through more research on the interactions between CTCs and their microenvironment. We account for all such mechanisms by defining a transition probability (which is itself a random variable) of a random walker dispersing from one site to another, thus creating a quantitative and computational framework for the seed-and-soil hypothesis as an ensemble based first step, then can be further refined primarily by using larger, better, and more targeted databases such as ones that focus on specific genotypes or phenotypes, or by more refined modeling of the correlations between the trapping of a CTC at a specific site, and the probability of secondary tumor growth at that location.

The Markov chain dynamical system takes place on a metastatic network based model of the disease, which we calculate based on the available data over large populations of patients.

Table 1.1: **Metastatic site numbering system**

| # | Name | # | Name |
|---|---|---|---|
| 1 | Adrenal* | 26 | Omentum* |
| 2 | Anus | 27 | Ovaries |
| 3 | Appendix | 28 | Pancreas* |
| 4 | Bile Duct | 29 | Penis |
| 5 | Bladder | 30 | Pericardium* |
| 6 | Bone* | 31 | Peritoneum* |
| 7 | Brain* | 32 | Pharynx |
| 8 | Branchial Cyst | 33 | Pleura* |
| 9 | Breast | 34 | Prostate* |
| 10 | Cervix | 35 | Rectum |
| 11 | Colon | 36 | Retroperitoneum |
| 12 | Diaphragm* | 37 | Salivary |
| 13 | Duodenum | 38 | Skeletal Muscle* |
| 14 | Esophagus | 39 | Skin* |
| 15 | Eye | 40 | Small Intestine* |
| 16 | Gallbladder* | 41 | Spleen* |
| 17 | Heart* | 42 | Stomach* |
| 18 | Kidney* | 43 | Testes |
| 19 | Large Intestine* | 44 | Thyroid* |
| 20 | Larynx | 45 | Tongue |
| 21 | Lip* | 46 | Tonsil |
| 22 | Liver* | 47 | Unknown |
| 23 | Lung* | 48 | Uterus* |
| 24 | Lymph Nodes (reg)* | 49 | Vagina* |
| 25 | Lymph Nodes (dist)* | 50 | Vulva |

Site numbering system used in transition matrix and network model. The * indicates an entry in the target vector associated with lung cancer primary from the dataset of [19].

In particular, we use the data described in the autopsy analysis in [19] in which metastatic distributions in a population of 3827 deceased cancer patients were analyzed. None of the patients received chemotherapy nor radiation. The autopsies were performed between 1914 and 1943 at 5 separate affiliated centers, with an ensemble distribution of 41 primary tumor types, and 30 metastatic locations. Figure 1.2 shows histograms of the number of metastases

found at the various sites in the population. Figure 1.2(a) shows the metastatic distribution in the entire population, while Figure 1.2(b) shows the distribution in the subset of the population with primary lung cancer. We note that this data offers no particular information on the time history of the disease for the population or for individual patients - only the long-time metastatic distribution in a population of patients, where long-time is associated with end of life, a timescale that varies significantly from patient to patient (even those with nominally the same disease). Although the initial analysis focuses on a model for primary lung cancer, the approach would work equally well for all of the main tumor types (which is outlined in Chapter 4).

Network based models of disease progression have been developed recently in various contexts such as the spread of computer viruses [4], general human diseases [27], and even cancer metastasis [10], but as far as we are aware, our Markov chain/random walk approach to modeling the dynamics of the disease on networks constructed for each primary cancer type from patient populations offers a new and potentially promising computational framework for simulating disease progression. More general developments on the structure and dynamics on networks can be found in the recent works [48, 50, 51, 52, 53, 74]. For brief introductions to some of the mathematical ideas developed in this study, see [18, 20, 25, 67].

The classic view of metastatic progression, framed in part by the seed-and-soil hypothesis of Paget [62], is that cancer spreads from the primary tumor site to distant metastatic locations in a unidirectional way. The seeds responsible for the spread are CTCs ([56, 77, 79]) that detach from the primary tumor, enter the bloodstream and lymphatic system [79], and

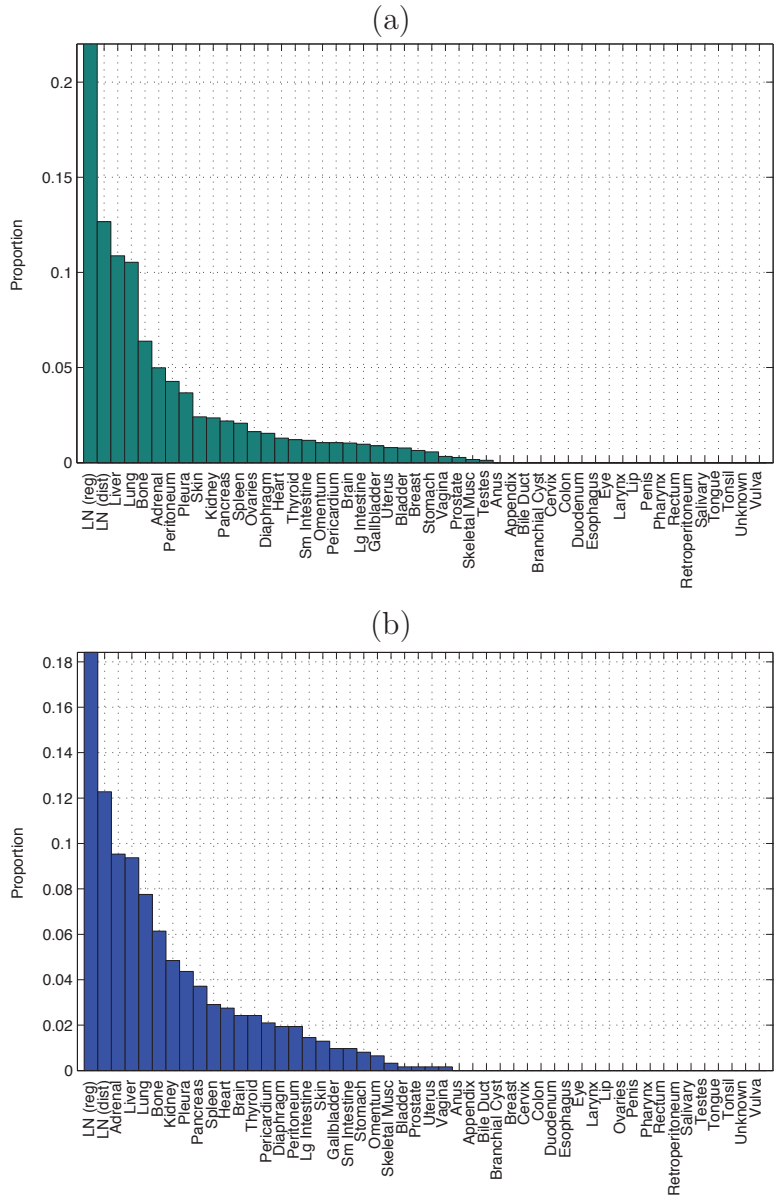


Figure 1.2: Metastatic distributions from autopsy data

Metastatic distributions from autopsy dataset extracted from 3827 patients [19]. Y-axis in each graph represents a proportion between 0 and 1. The sum of all the heights is 1. These are the two key probability distributions used to 'train' our lung cancer progression model. (a) Overall metastatic distribution including all primaries. We call this distribution the 'all' cancer distribution as it includes all primary cancer types.; (b) Distribution of metastases associated with primary lung cancer. We call this distribution the 'target' distribution that we label $\vec{v}_T$.

travel to new distant locations. If conditions are favorable, this initiates a complex [9, 23, 78] and not well understood metastatic cascade, ultimately leading to tumor growth at distant anatomic sites if their soil is hospitable [62]. The exclusively unidirectional nature of this process has been challenged recently in a series of articles [2, 11, 41, 44, 60, 68], which use mouse models to show a mechanism by which CTCs from the primary tumor can re-enter the primary, a process called 'self-seeding' [60]. These authors further comment that "it is tempting to speculate that self-seeding might occur not only at the primary tumor site, but also at distinct metastatic sites, ... each site being a nesting ground." The possibility of metastasis from metastases has also been discussed in [6, 33]. While the underlying 'agent' responsible for the spread of cancer is the CTC, the disease progression pathways in different patients can be both predictable (from a statistical viewpoint), but often unpredictable and surprisingly distinct in patients with nominally the same disease [21, 81], prompting the question "how can metastatic pathways be predictable and unpredictable at the same time [11]?"

Motivated in part by these questions, we develop our model for cancer progression and use it to identify and quantify the multidirectional pathways and timescales associated with metastatic spread for primary lung cancer.

While stochastic in nature, our model shows that a defining aspect of both pathway selection and timescale determination is whether the disease spreads from the primary tumor to a metastatic site that is either a 'spreader' (adrenal gland and kidney) or a 'sponge' (regional lymph nodes, liver, bone). In contrast to the traditional view of cancer metastasis as

a unidirectional process starting at the primary site and spreading to distant sites as time progresses, our model supports and quantifies the view that there are important multidirectional aspects to metastatic progression. These fall under 3 general classes: (i) self-seeding of the primary tumor, (ii) reseeding of the primary tumor from a metastatic site (primary reseeding), and (iii) reseeding of metastatic tumors (metastasis reseeding).

Using a discrete Markov chain [58] system of equations applied to a large autopsy dataset of untreated patients with cancer [19], we quantify the likelihoods of the top metastatic pathways in terms of probabilities and conduct Monte Carlo computer simulations of cancer progression that statistically reflect the autopsy data about a (non-Gaussian) distribution of disease. The stochastic Markov chain dynamical system takes place on a metastatic network-based model of disease progression that we construct based on available autopsy data over large populations of patients. To obtain our baseline model, we use the data described in an autopsy analysis [19] in which metastatic tumor distributions in a population of 3827 untreated deceased cancer patients were recorded; 163 of these had primary lung cancer of some type, distributing a total of 619 metastatic tumors across 27 different sites. Information on lung cancer type in this dataset is not possible to obtain as the samples were collected before the widespread use of immunohistochemistry (1914-1943), without which, the subcategorization of non-small cell lung cancer is unreliable. However, it is probably safe to assume that the distribution of lung cancer type was not significantly different than current distributions, roughly 40% adenocarcinoma, 30% squamous cell carcinoma, 9% large cell carcinoma, and 21% small cell carcinoma.

## 1.3 Tumor entropy

Metastatic cancer is a dynamic disease of relentlessly increasing entropy. From an initial primary tumor located at a single anatomical site, the metastatic cascade leads to a proliferation of tumors at other sites on a timescale of months, or years in most cases, if left untreated [9, 23, 77, 78]. Entropy is a quantity associated with notions of complexity and predictability used primarily in two distinct fields: information theory [13, 37, 71] and statistical thermodynamics [40]. It is used to quantify the level of 'disorder' associated with a dynamical process that has a potentially large number of sites that it can occupy [16, 65]. Systems that can visit a large number of sites with relatively equal probability have higher entropy (they are considered more disordered and less predictable) than systems that can only occupy a few sites with widely separated probabilities (considered less disordered and more predictable). Our goal in this study is to demonstrate how the notion of entropy can be used in the context of metastatic spread to quantify and compare the complexity of the 12 most prevalent cancer types worldwide.

To fix ideas further, suppose each anatomical site where a primary or metastatic tumor could appear is indexed by '$i'$, $(i = 1, ...N)$. Let $\sigma_i$ represent the probability that site 'i' is occupied (i.e. has a metastatic tumor), and let $\vec{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N) \in R^N$ represent a probability mass distribution over a collection of potentially occupied sites, so that $\sum_{i=1}^{N} \sigma_i = 1$, with $0 \leq \sigma_i \leq 1$. The level of disorder associated with the distribution $\vec{\sigma}$ is captured by a scalar quantity $H_N$, called the 'entropy' of the state $\vec{\sigma}$. As explained later, it is a quantity that is a function both of N, and the way the probabilities are distributed among the N sites.

The lowest entropy state, corresponding to the one of least disorder, would be represented by a distribution such as $\vec{\sigma} = (0, 0, 1, 0, 0, \ldots, 0)$, in which case $H_N = 0$. In this distribution, state $i = 3$ is occupied with probability 1, making it predictively certain. Typically, this site would be the anatomical location of the primary tumor in a patient whose disease has not yet progressed to other sites. The highest entropy state, corresponding to the one of most disorder, would be represented by the 'uniform' distribution $\vec{\sigma} = (\frac{1}{N}, \frac{1}{N}, \frac{1}{N}, \frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N})$. For this uniform distribution, each site is occupied with equal probability. This distribution, which constitutes an upper bound on the entropy, represents a state of maximal disorder, corresponding to the least predictable state. The point we want to emphasize is that associated with any specific probabilistic distribution of occupied sites (typically falling between the above two extremes), is a quantitative notion of 'disorder', which in turn is related to the system's predictability and complexity [16, 65]. Since each cancer type has a different empirical metastatic tumor distribution, each will have a different metastatic entropy value and these entropy values can be thought of as convenient 'surrogates' representing metastatic complexity and disorder.

Chapter 2

# A stochastic Markov chain model to describe cancer metastasis

## 2.1 Introduction

This section focuses on building a Markov chain model for primary lung cancer and highlighting some preliminary results. An initial guess of the model is constructed from the data provided in [19] and then iterated on until a final model is reached. We then use the entries of the transition matrix and it's dynamics to classify the metastatic sites of the model and to translate the results into easily understood and useful information.

## 2.2 Methods

Because we are computing the entries of a $50 \times 50$ matrix using only the 50 entries of our target steady-state, the solution to this problem is not unique, a problem which is addressed in the works of [17, 30, 31] for example. In those papers, the solution to this

constrained linear inverse problem is obtained by identifying the transition matrix that satisfies a certain maximum entropy condition, and also one obtained by satisfying a least-squares condition. More relevant to our problem is a criterion which targets a family of solutions by pre-conditioning the search on an approximate transition matrix informed by the data, followed by an iteration process which then adjusts the entries until a transition matrix with the correct steady-state is obtained. We show that this process converges, and we use the algorithm to create an ensemble of transition matrices whose entries are best interpreted as (approximately) normally distributed random variables. We then characterize the ensemble of stochastic transition matrices using the means and variances of the singular value distributions [28] associated with the ensemble.

### 2.2.1 Algorithm to compute the Markov transition matrix

The three key steps in computing the transition matrix are:

(i) Step 1 - The choice of initial matrix $A_0$: First, an approximate transition matrix, $A_0$, is obtained based on information we extract directly from the data set [19]. For the 'lung row' of $A_0$, we use the lung target distribution shown in Figure 1.2(b), which is the metastatic distribution in a population of people with lung cancer primary tumors. This is our first approximation to how the outgoing edges from the lung are weighted. On all of the other 49 rows, we use the all cancer distribution shown in Figure 1.2(a). Since we do not know, a priori, how any of the other metastatic sites communicate with any of the others, we use this 'agnostic' distribution for all of these non-lung rows. Two key properties of $A_0$ constructed

this way are that it has Rank = 2 (i.e. only two linearly independent rows), and it does not have our target distribution shown in Figure 1.2(b) as a steady-state, hence we know $A_0$ is not the correct transition matrix for lung cancer. Therefore, we perform an iteration process in Step 2 which adjusts the entries of $A_0$ to arrive at a final transition matrix $A_f$ that has higher rank (typically the same rank as the number of entries in the target vector), and has the target distribution (Figure 1.2(b)) as a steady-state.

(ii) Step 2 - The iteration process to $A_f$: $A_0$, is then used to start an iteration process where the entries are adjusted iteratively, using randomized adjustments, until its steady-state distribution converges to the target distribution. The converged matrix obtained after this process is what we call the 'trained' lung cancer matrix, $A_f$ . We will discuss this key step further below.

(iii) Step 3 - Creating an ensemble of $A_f$'s: Because the iterative procedure is based on random adjustments of the matrix entries, and because we adjust the entries only up to some pre-determined numerical value defined as our convergence threshold (typically chosen to be $O(10^{-5})$), the transition matrices produced from Step 2 should be thought of as having entries that have some inherent probability distribution associated with them, with a sample mean and variance obtained by collecting an ensemble of these matrices. We will show two of the key edge probability distributions (lung to regional lymph nodes, and lung to adrenal) and also discuss the statistical spread of the ensemble of transition matrices using their singular value distributions as a diagnostic tool.

### 2.2.2 Convergence of the algorithm

We now describe Step 2 of our algorithm in more detail, the iterative training stage which takes us from our initial matrix $A_0$, to our final matrix $A_f$. Define the transition matrix after step $j$ in the iteration process to be $A_j$, with corresponding steady-state $\vec{v}_\infty^{(j)}$ defined as

$$\vec{v}_\infty^{(j)}(A_j - \mathbf{I}) = 0. \tag{2.1}$$

Our goal is to find the entries of $A_j$ so that

$$\vec{v}_T(A_j - \mathbf{I}) = 0. \tag{2.2}$$

i.e. so that $\|\vec{v}_\infty^{(j)} - \vec{v}_T\|^2 = 0$. We do this iteratively as follows. Since $\vec{v}_T \neq \vec{v}_\infty^{(j)}$, we can define a 'residual' at step $j$:

$$\vec{v}_T(A_j - \mathbf{I}) = \vec{r}_j \equiv (\vec{v}_T - \vec{v}_\infty^{(j)})(A_j - \mathbf{I}), \tag{2.3}$$

where $\|\vec{r}_j\|^2 \neq 0$. Our goal is to find the entries of $A_j$ so that $\|\vec{r}_j\|^2 \leq \epsilon << 1$, where $\epsilon$ is defined as our numerical convergence threshold. In practice, we do this by calculating $\|\vec{v}_T - \vec{v}_\infty^{(j)}\|^2$ directly and iterate the entries of $A_j$ until $\|\vec{v}_T - \vec{v}_\infty^{(j)}\|^2 < \epsilon$, where typically we

take $\epsilon = O(10^{-5})$. Stated more generally, our goal is to solve the following linear constrained optimization problem. Given a target vector $\vec{v}_T$, find the entries $a_{ij}$ of the matrix A to minimize the Euclidean norm of the residual vector $\vec{r}$, where:

$$\vec{v}_T(A - \mathbf{I}) = \vec{r}. \tag{2.4}$$

The constraints are $0 \leq a_{ij} \leq 1$, and $\sum_{j=1}^{50} a_{ij} = 1$. Most importantly, we have preconditioned the iterative process in Step 1 on our particular initial matrix $A_0$. The general framing of this problem as a constrained optimization problem is discussed in [17, 30, 31].

To do this, we iteratively adjust the entries of $A_j$ at each step (so as to maintain the constraint that all rows sum to one) according to the following algorithm:

1. Calculate the residual $\vec{r}_j$ at step $j$, starting with $A_0, (j = 0)$;
2. Pick the column of $A_j$ corresponding to the maximum entry of $\vec{r}_j$;
3. Pick the column of $A_j$ corresponding to the minimum entry of $\vec{r}_j$;
4. Pick a row of $A_j$ at random;
5. Decrease the entry of $A_j$ selected in step (ii) by $\delta$, increase the entry of $A_j$ selected in step (iii) by $\delta$, where $\delta$ is scaled with the size of $\|\vec{r}_j\|^2$. This new matrix is $A_{j+1}$;
6. Calculate the new $\|\vec{r}_{j+1}\|^2$ and stop if $\|\vec{r}_{j+1}\|^2 < \epsilon$. Otherwise go to step (ii) and repeat the process.

Because of the randomized nature of the algorithm, and because of the finite threshold of convergence, the converged final matrix $A_f$ will be slightly different each time the iterative process is carried out, even when all the trained matrices start with the same initial $A_0$. Thus, we carry out the iteration and convergence process, producing an ensemble of 1000 final transition matrices $A_f$ , and we show the convergence (down to $O(10^{-5})$) of the ensemble in Figure 2.1 (plotted on a semi-log plot). The solid curve is the average convergence rate computed from the 1000 training sessions, while the error bars show the standard deviations associated with the ensemble, showing the spread of the convergence rates, which are relatively tight.

### 2.2.3 Singular values and properties of the ensemble

A very useful diagnostic tool to characterize the structure and understand the statistical spread associated with the matrices in the ensemble are the singular values, $\lambda_n(\lambda_1 > \lambda_2 > \ldots > \lambda_27 > 0)$, associated with the collection of $A_f$'s. These are shown in Figure 2.2, plotted from largest to smallest. Values shown (as open circles) are the sample means associated with the singular values of the ensemble of 1000 converged matrices $A_f$, all trained using the same initial matrix $A_0$. The error bars show the sample standard deviations, which are small. The 27 non-zero singular values reflect the fact that there are 27 entries in the steady-state distribution for primary lung cancer. An equivalent way to say this is that the rank of $A_f$ is 27, while the nullspace dimension is (approximately) 23. The standard deviations show the statistical spread associated with two sources of uncertainty, one is the random

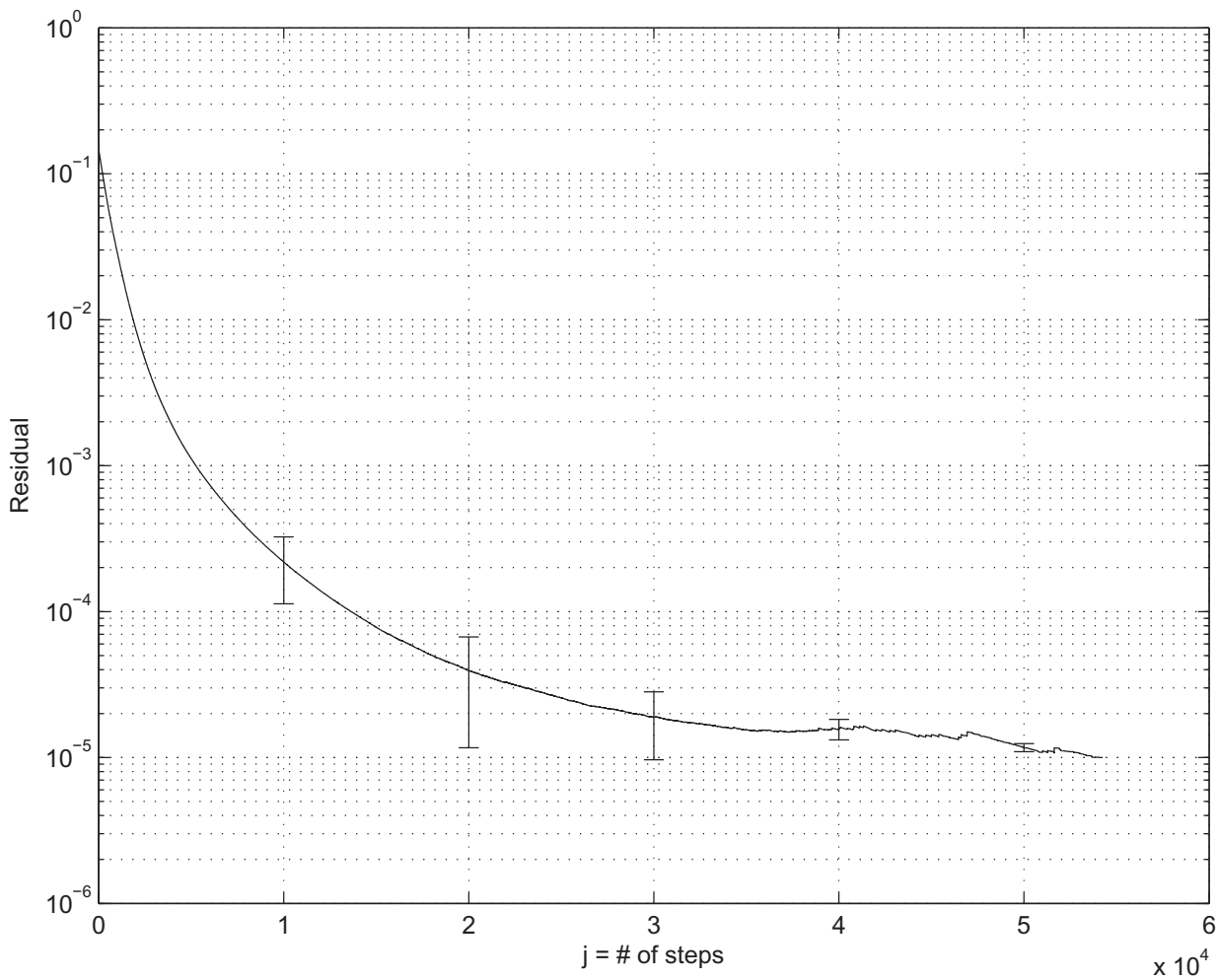


Figure 2.1: Ensemble convergence graph
Ensemble convergence to $A_f$, starting from $A_0$. y-axis is $\|\vec{r}_j\|^2$, x-axis is step $j$. We use an ensemble of 1000 trained matrices $A_f$, each conditioned on the same initial matrix $A_0$. The average convergence curve is shown, along with standard deviations marked along each decade showing the spread associated with the convergence rates.

search algorithm we use to obtain convergence, and the other is the convergence threshold, which we typically take to be $O(10^{-5})$. The small standard deviations indicate that the algorithm is converging to the same final $A_f$ , within a relatively small range of statistical spread. On this graph, we also show the least squares curve fit to singular values $\lambda_4$ through

$\lambda_{24}$, which follow a slope $\beta \sim -0.1389$, indicating that the singular values roughly decrease like $\lambda_n \sim \alpha \exp(-\beta_n)(\alpha \sim 0.1901)$. The two diamond shaped data points on the graph correspond to the two singular values of $A_0$ reflecting the linear independence of the two distributions from Figure 1.2 that we use in $A_0$. We point out that the $A_f$'s should not be viewed as small perturbations of $A_0$ - our convergence algorithm starts with a rank 2 matrix and generates an ensemble of (approximately) rank 27 matrices all within a relatively tight statistical spread.

We also show one other set of singular values on the graph with the asterix data points. To test the robustness of the ensemble with respect to perturbations of the initial matrix $A_0$, we start the search with an initial matrix of the form $A_0 + \epsilon A_1$. Here, the perturbation matrix $A_1$ is a $50 \times 50$ rank 2 matrix obtained by giving each entry in the lung row a uniformly distributed random number in the interval [-1,1], and each entry in all the other rows another uniformly distributed random number in the interval [-1,1]. This creates a random rank 2 matrix. The perturbation parameter $\epsilon$ is chosen so that the perturbation size is (roughly) 5% as compared with the average row value of $A_0$. The asterix data points, which correspond to a converged $A_f$ below a threshold of $O(10^{-10})$, all fall within the one standard deviation bars of the unperturbed values, again showing that the final converged matrix is relatively robust to small changes in the initial matrix $A_0$. For definiteness, when we make conclusions associated with Monte Carlo simulations, we use the ensemble averaged set of $A_f$'s obtained over a set of 1000 converged matrices, each converged to within $O(10^{-5})$. Because of this,

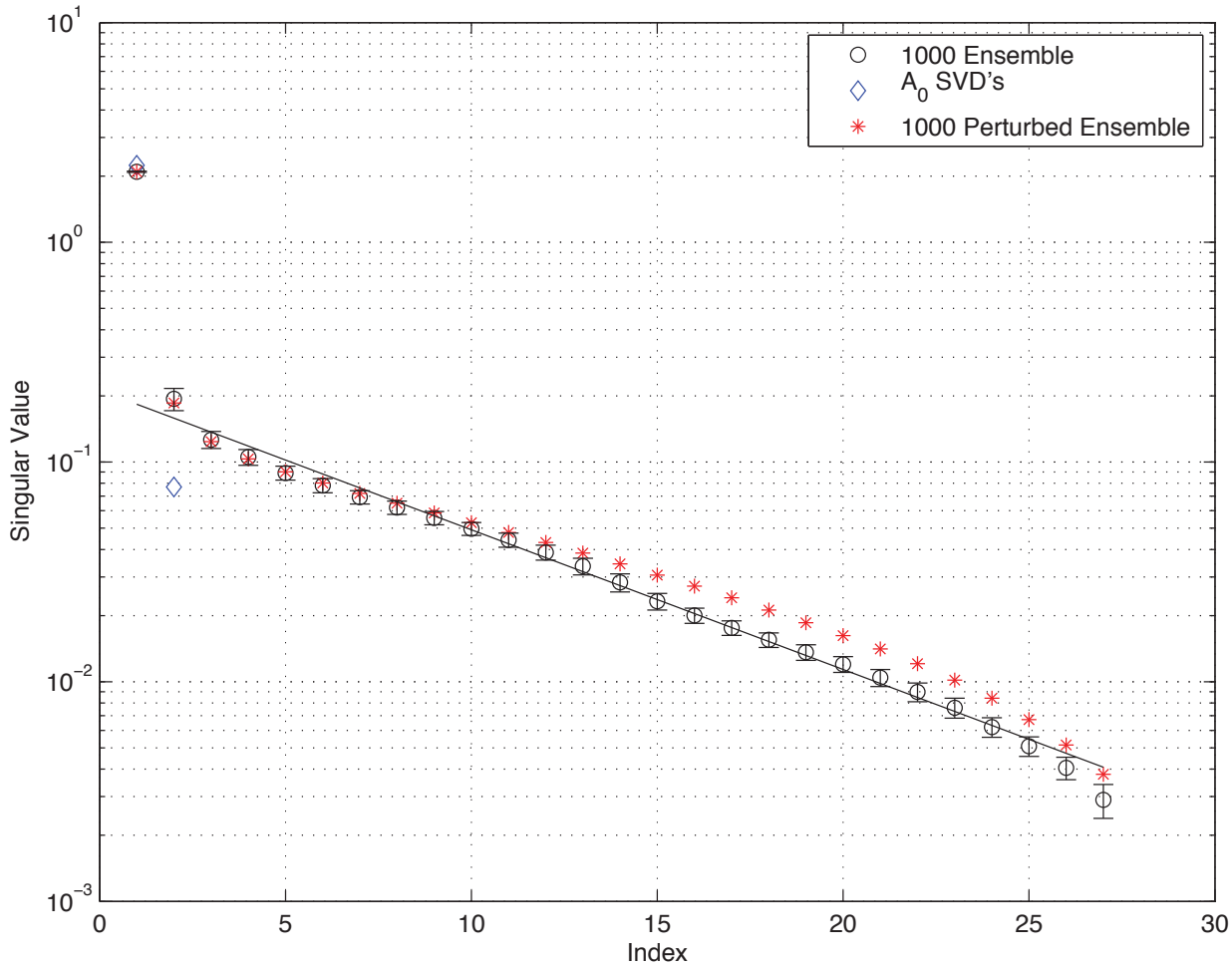


Figure 2.2: Singular value distribution

Average distribution of the 27 non-zero singular values associated with the ensemble of 1000 matrices $A_f$ all obtained using the same $A_0$. $x$-axis is the index $n$, $y$-axis is $\lambda_n$. Data points (open circles) indicate the sample average, with error bars showing the sample standard deviations. Line is a least squares curve fit through $\lambda_4$ through $\lambda_{24}$, showing linear decrease with exponent $\beta = -0.1389$. The 27 non-zero singular values reflect the fact that there are 27 entries in the steady-state target distribution for primary lung cancer. The two diamond shaped data points are the two singular values associated with the initial matrix $A_0$. The 27 'asterix' data points are those obtained from a trained matrix using a perturbed $A_0$, with Rank 2 perturbation. See text for details.

we view the transition probabilities of the Markov chain, i.e. the edge values in our network,

as themselves being random variables, with a standard deviation that we can characterize.

## 2.3 Results

In this section we describe three main results from the model. First, the model separates the 27 non-zero sites from Figure 1.2(b) into what we call 'first-order' sites (20 of these), and 'second-order' sites (7 of these). Second, the model quantifies the ability of each site to self-seed by ranking the average edge weight of each site back to itself (see [60]). Of these, the strongest self-seeders are the lymph nodes, bone, kidney, and lung. Third, the model allows us to calculate a time-ordering (model based) associated with metastatic progression. This is achieved by performing Monte Carlo simulations of the mean first-passage times from the lung site to each of the other sites in the network. The mean first-passage time is the average number of edges a random walker must traverse in order to hit a given site, hence the number is not restricted to take on discrete integer values. We think of these mean first-passage times as the proxy timescale for progression. In principle, they can be calculated analytically using the fundamental matrix (see [29]), but in practice, since this involves inverting the $50 \times 50$ transition matrix, it is far more convenient to obtain the results numerically via Monte Carlo simulations. The results will be described in terms of a 'random walker' leaving the lung site and traversing the network, moving from site to site along one of the outgoing edges available to it at the site it is leaving, choosing a given edge with the probability corresponding to its weighting.

### 2.3.1 Description of the Markov chain model

With the stochastic transition matrix $A_f$, we briefly describe the basic features and interpretations of a Markov dynamical system model which we write as:

$$\vec{v}_{k+1} \quad = \quad \vec{v}_k A_f, \quad (k = 0, 1, 2, \dots) \tag{2.5}$$

The matrix $A_f$ is our transition matrix which is applied to a state-vector $\vec{v}_k$ at each discrete time-step $k$ to advance to step $k+1$. Thus, it is easy to see that:

$$\vec{v}_k = \vec{v}_0 A_f^k \tag{2.6}$$

where $\vec{v}_0$ is the initial-state vector. The underlying dynamics associated with disease progression is interpreted as a random walk on the weighted directed network defined by the entries of the transition matrix.

### 2.3.2 The state vectors and definition of the steady-state

To interpret the meaning of the initial-state vector and the transition matrix, one should think of the patient's initial tumor distribution in terms of probabilities, or 'uncertainties'. Thus, an initial-state vector with a 1 in the 23rd entry:

$$\vec{v}_0 = (0, 0, 0, 0, 0, 0, 0, \dots, 1, \dots)$$

in our 50 node model indicates, with absolute certainty, that the patient has a primary tumor located in the 'lung' (position 23). At the other extreme, we may have an initial-state vector:

$$\vec{v}_0 = (1/50, 1/50, 1/50, 1/50, 1/50, 1/50, \dots)$$

which indicates that all locations of the initial tumor distribution are equally likely. One interpretation of this is that we have no information at all about where the primary tumor is located. A third possibility is that we have *some* limited information about the initial tumor distribution, but not completely certain information, thus an initial-state vector:

$$\vec{v}_0 = (1/2, 0, 0, 0, 0, 0, 1/2, 0, 0, 0, 0, 0, 0, 0, 0, 0, \dots)$$

would indicate that we think it likely that there is a primary tumor in the 'adrenal' (position 1) or 'brain' (position 7), but we are not sure which.

Then, we can ask how this initial information propagates forward in time as the disease progresses. To advance one-step forward in time, we apply the transition matrix once to the initial-state vector, thus:

$$\vec{v}_1 = \vec{v}_0 A_f.$$

This gives us our new state-vector $\vec{v}_1$ after step one. For the next step, we apply the transition matrix again, this time to $\vec{v}_1$:

$$\vec{v}_2 = \vec{v}_1 A_f = \vec{v}_0 A_f^2.$$

The dynamical system proceeds according to Equation (2.6) in a manner consistent with the schematic diagram from Figure 1.1. As described in the introduction, it is best to think of the entries of the state-vector as probabilities for metastases developing at each of the discrete sites in our model (and in the data set), thus for the seed to take root in the soil. The entries of the state-vector $\vec{v}_k$ continually get redistributed in time, as measured in discrete steps $k$, until they reach the target steady-state distribution. A different interpretation of the entries of the state-vector at each discrete step is that they reflect the *ensemble statistical distribution* of a collection of agents executing a random walk across the network. We should point out, however, that for the ensemble of random-walkers all leaving from the lung site, the best way to measure the passage of time is via *mean first-passage times* to each of the sites, which we compute using Monte Carlo simulations. It is important to keep in mind that since the transition matrix is constructed based on an *autopsy* data set, there is no direct information available on time-histories of progression, only tumor distribution at time of death. A big advantage of using this data set is that we are able to build a model based on the 'natural' progression of the disease (i.e. untreated patients), whereas clinical data on time-histories of progression for untreated patients do not exist, as far as we are aware. Therefore, our challenge is to extract as much information as we can using the autopsy data

set [19], keeping in mind that time should be interpreted only as the model timescale of progression.

Now comes a natural and important question. After long-times ($k$ large), is there some steady-state distribution that is achieved by the model? Correspondingly, given a particular primary tumor, what are long-term probabilistic distributions of possible metastases? We call this distribution vector $\vec{v}_\infty^{(0)}$, and define it as:

$$\vec{v}_\infty^{(0)} = \lim_{k\to\infty} \vec{v}_0 A_f^k. \tag{2.7}$$

Notice that if a steady-state distribution is achieved, then for sufficiently large $k$, $\vec{v}_{k+1}^{(0)} \sim \vec{v}_k^{(0)}$, and since

$$\vec{v}_{k+1}^{(0)} = \vec{v}_k^{(0)} A_f, \tag{2.8}$$

this implies that

$$\vec{v}_\infty^{(0)} = \vec{v}_\infty^{(0)} A_f. \tag{2.9}$$

Thus

$$\vec{v}_\infty^{(0)}(A_f - \mathbf{I}) = 0, \tag{2.10}$$

which means that $\vec{v}_\infty^{(0)}$ is a left-eigenvector of $A_f$ corresponding to eigenvalue $\lambda = 1$. This is a crucial and practical observation that allows us to calculate the steady-state distribution $\vec{v}_\infty^{(0)}$ directly from the transition matrix. Since the rows of $A_f$ add to one, it always has at least one eigenvalue that is 1, hence there is always at least one steady-state distribution, but there may be more than one — this depends in detail on the matrix structure, something the eigenvalue distribution [28] can reveal.

The target distribution for lung cancer shown in Figure 1.2(b) and labeled $\vec{v}_T$ is not a steady-state for the matrix $A_0$, i.e.

$$\vec{v}_T(A_0 - I) = (\vec{v}_T - \vec{v}_\infty^{(0)})(A_0 - I) \neq 0, \tag{2.11}$$

since $\|\vec{v}_T - \vec{v}_\infty^{(0)}\|^2 \neq 0$.

### 2.3.3 Structure of the lung cancer matrix and convergence to the steady-state

Figure 2.3 shows the network diagram associated with the ensemble averaged converged matrix - this is the lung cancer network conditioned on our initial guess $A_0$ averaged over 1000 training sessions. Each of the sites has incoming and outgoing edges (denoted with arrow heads) which connect it to other sites in the target distribution where the cancer can spread, and each of the edges have a probabilistic weighting (not shown), with the constraint that the weightings associated with all outgoing edges at each site must sum to 1. The disease

spreads across the network from an initial site following a random walk. To minimize the number of edges depicted in the figure, we have combined incoming and outgoing edges whenever possible, and placed arrow heads on both ends of an edge, instead of plotting the two edges separately.

In Figure 2.4 we plot the (mean) edge weightings of the outgoing edges from the lung, as compared with the values of the target distribution shown in Figure 1.2(b). The differences show that the values in the lung row of $A_f$ have adjusted from their initial values in $A_0$. Figure 2.5 and Figure 2.6 highlight our interpretation of the transition probabilities, or edge values of the network, as random variables. We show in these figures the distributions associated with the ensemble of lung to regional lymph node (Figure 2.5) edge values, and those associated with the lung to adrenal (Figure 2.6) edge values. In each case, we histogram the edge values from the 1000 converged matrices, and use the sample means and variances to overlay a corresponding normal distribution. The vertical dashed lines in Figures 2.5 and 2.6 show the initial value of the transition probability from lung to regional lymph nodes (Figure 2.5) and lung to adrenal (Figure 2.6). These initial values used in the matrix $A_0$ are obtained using the entire data set of DiSibio and French [19], i.e. over all primary cancer types. The converged Gaussian distributions shown in these figures, however, are specific to lung cancer only. The fact that the mean is clearly shifted to the left of the vertical line in Figure 2.5 indicates that the lung to regional lymph node connection for lung cancer is less significant, statistically, than for other cancer types. A possible anatomical explanation for this left shift could be the fact that regional lymph nodes, for lung cancer, are located

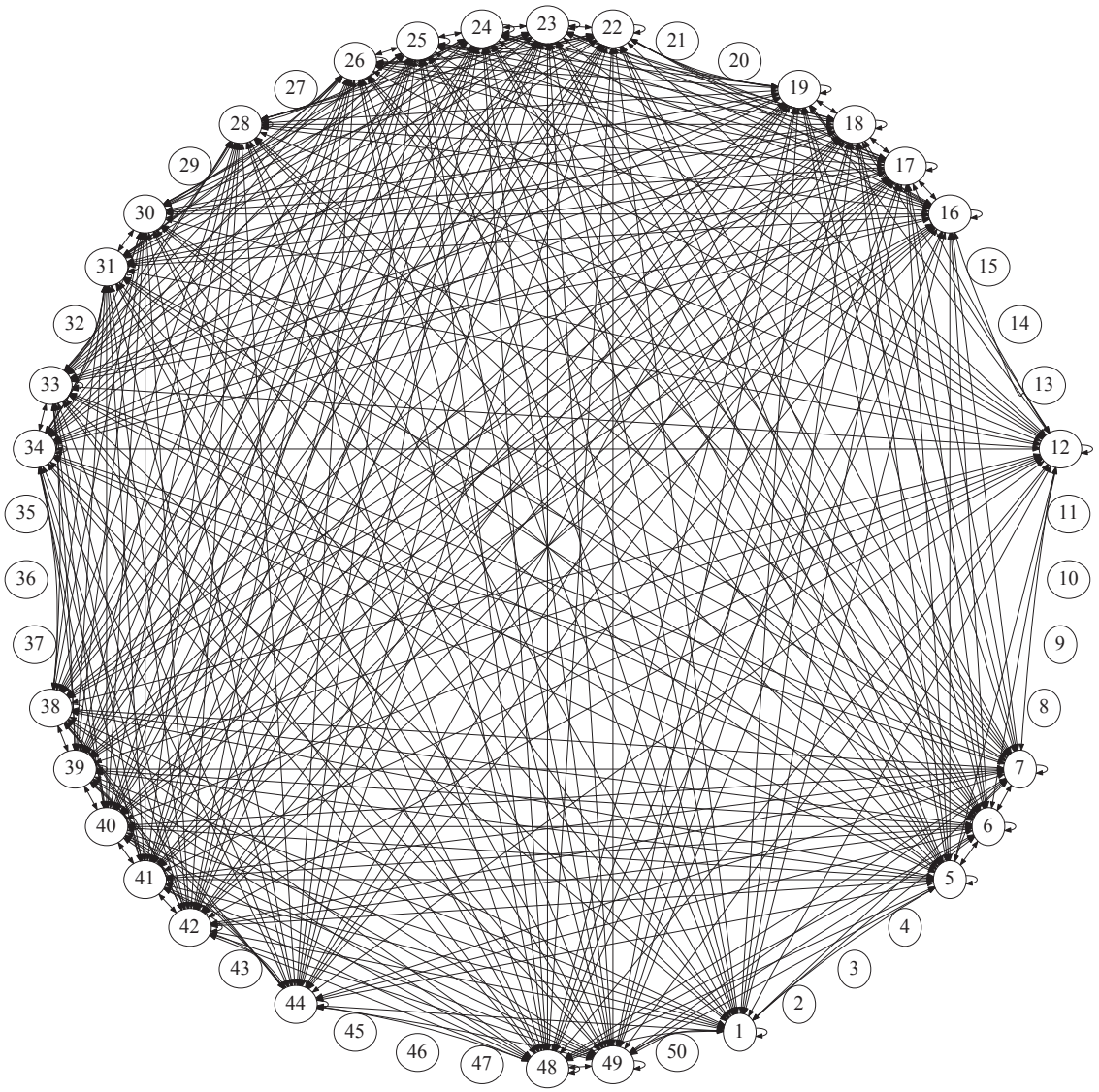


Figure 2.3: Converged lung cancer network

The converged lung cancer network shown as a circular, bi-directional, weighted graph. We use sample mean values for all edges connecting sites in the target distribution. The disease progresses from site 23 (lung) as a 'random walker' on this network. Arrow heads placed on the end or ends of the edges denote the direction of the connections. Edge weightings are not shown. There are 50 sites (defined in Table 1.1) obtained from the full data set of [19], with 'Lung' corresponding to site 23 placed on top. The 27 sites that are connected by edges are those from the target vector for lung cancer defined in Table 1.1.

very close to the lung itself, compared with their typical distance away from other primary tumor locations. Because of this unusually close proximity, regional lymph nodes could easily have been mistakingly considered as part of the lung in some of the autopsies in the series, effectively reducing the significance of the lung to regional lymph node connection. By contrast, the right shift of the mean, shown in Figure 2.6 for the lung to adrenal connection, would indicate that the lung to adrenal connection is statistically more important for lung cancer than for other primary cancer types. This could be due to the documented anatomic connection between lung and adrenal that is known, but has not, to date, been a particular focus of lung cancer metastasis studies.

The dynamical system defined by the Markov process:

$$\vec{v}_{k+1} \quad = \quad \vec{v}_k A_f, \quad (k = 0, 1, 2, , , , ) \tag{2.12}$$

can be thought of as governing the statistical distribution associated with random walkers traversing the network. Figures 2.7 and 2.8 show the dynamical progression of the initial state-vector, starting with an initial state-vector corresponding to a lung tumor, i.e. 1 in position 23, with 0's elsewhere. In the sequence, the target vector $\vec{v}_T$ is depicted with filled bars, while the vector $\vec{v}_k$ (for $k = 0, 1, 2, \infty$) is depicted with unfilled bars. Convergence to the target is exponential. By $k = 2$, convergence to the steady-state is essentially complete.

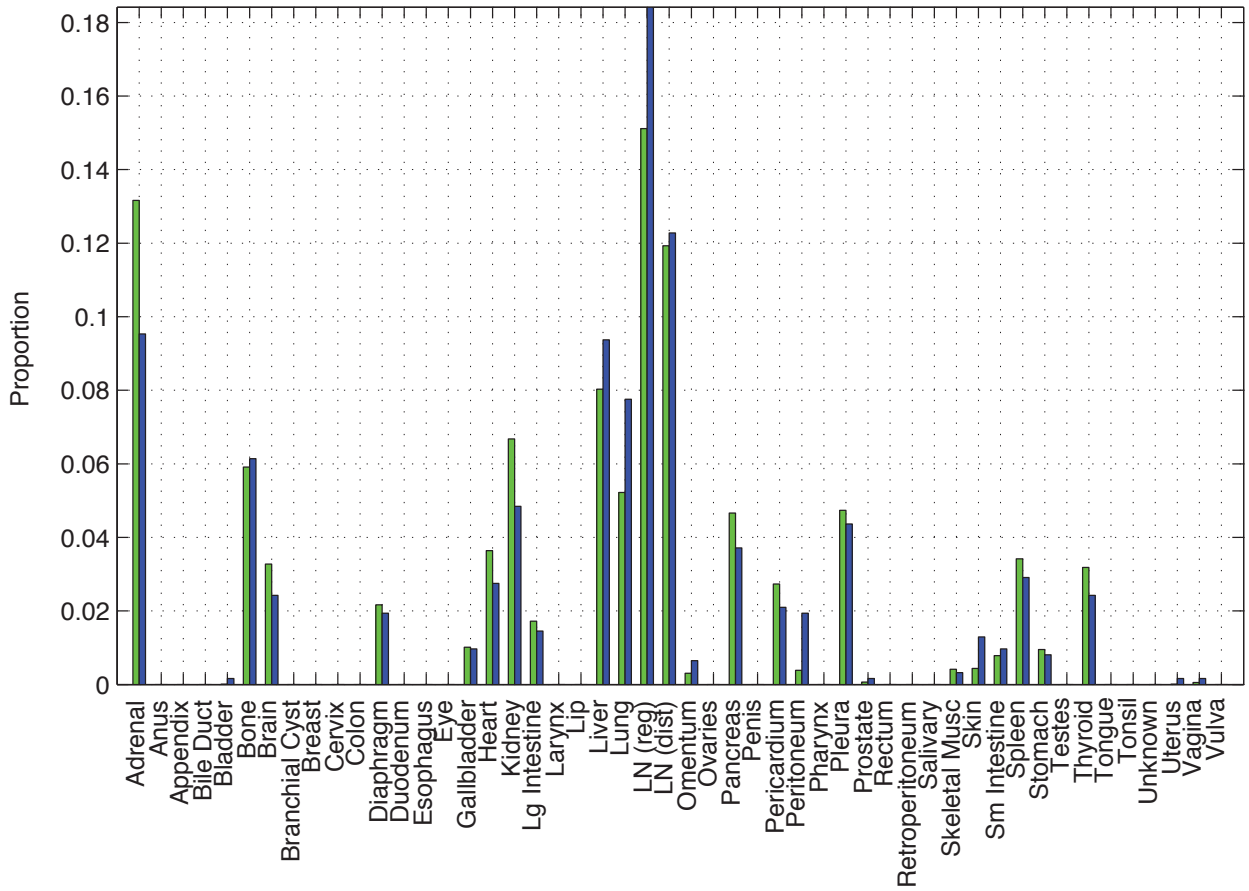


Figure 2.4: Outgoing edges from lung vs. target distribution
Weight of outgoing edges from the lung (using sample mean values from ensemble) as compared with the 'target' distribution.

### 2.3.4 First- and second-order sites

The 27 metastatic sites associated with lung cancer shown in the distribution of Figure 1.2(b) can be separated into two distinct groups in light of the ensemble averaged transition probabilities listed in decreasing order in Table 2.1. The middle column of this table shows the transition probability going directly from the lung to each of the 27 sites of the target vector (ensemble averaged $\pm$ standard deviations). The right column of the table shows

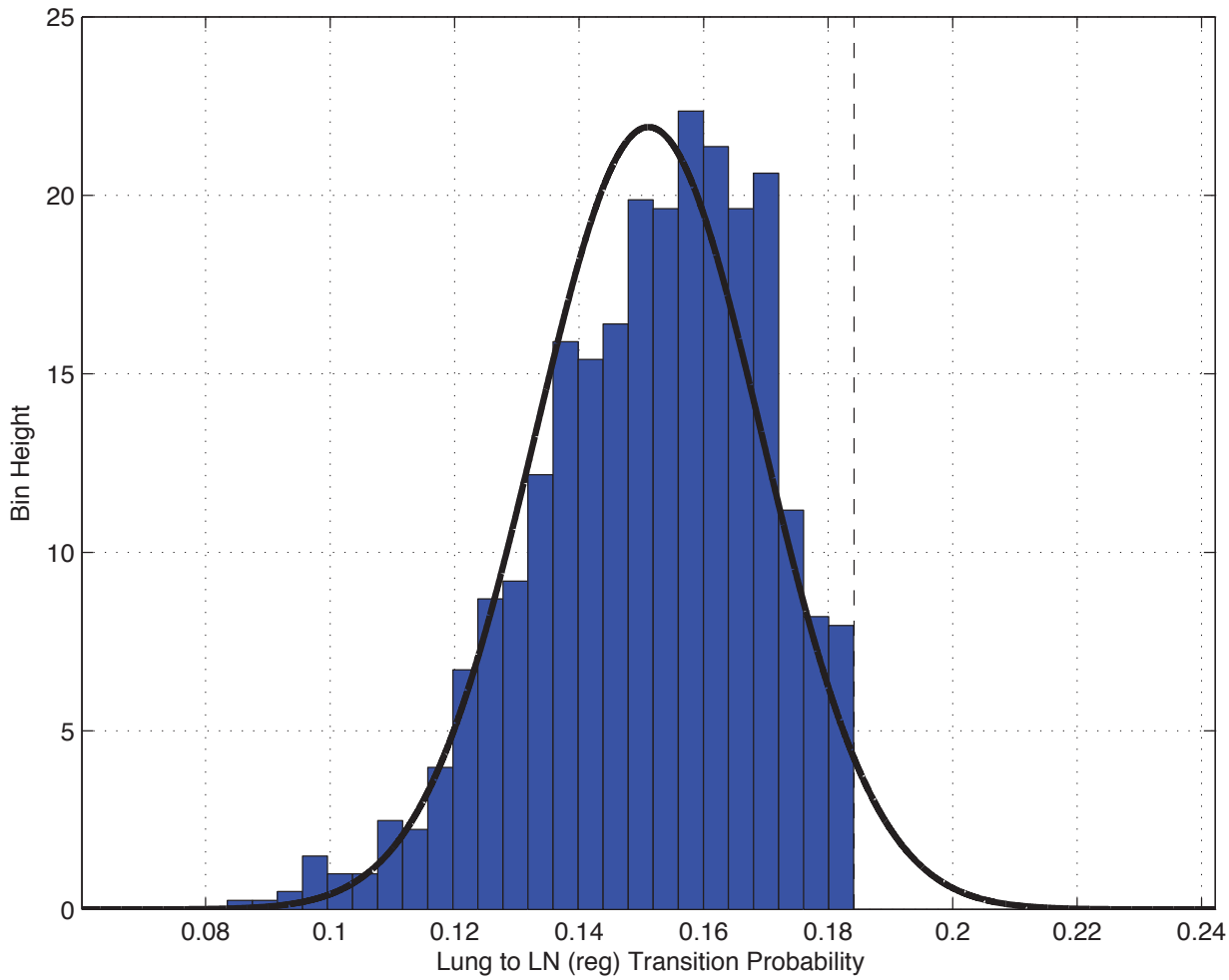


Figure 2.5: Lung → LN (reg) histogram
Histogram of edge values from lung to regional lymph nodes for 1000 trained $A_f$'s, showing that edge values (transition probabilities) are best thought of as random variables which are (approximately) normally distributed. Dashed vertical line shows initial edge value associated with $A_0$. Normal distribution with sample mean (0.15115) and variance (0.01821) is shown as overlay.

the most likely two-step path from lung to each of the sites listed on the left, via the most probable intermediate site. Thus it shows the product of the direct transition probability from lung to an intermediate site (in parentheses on right), times the transition probability from that intermediate site to the site listed on the left. When one compares these values

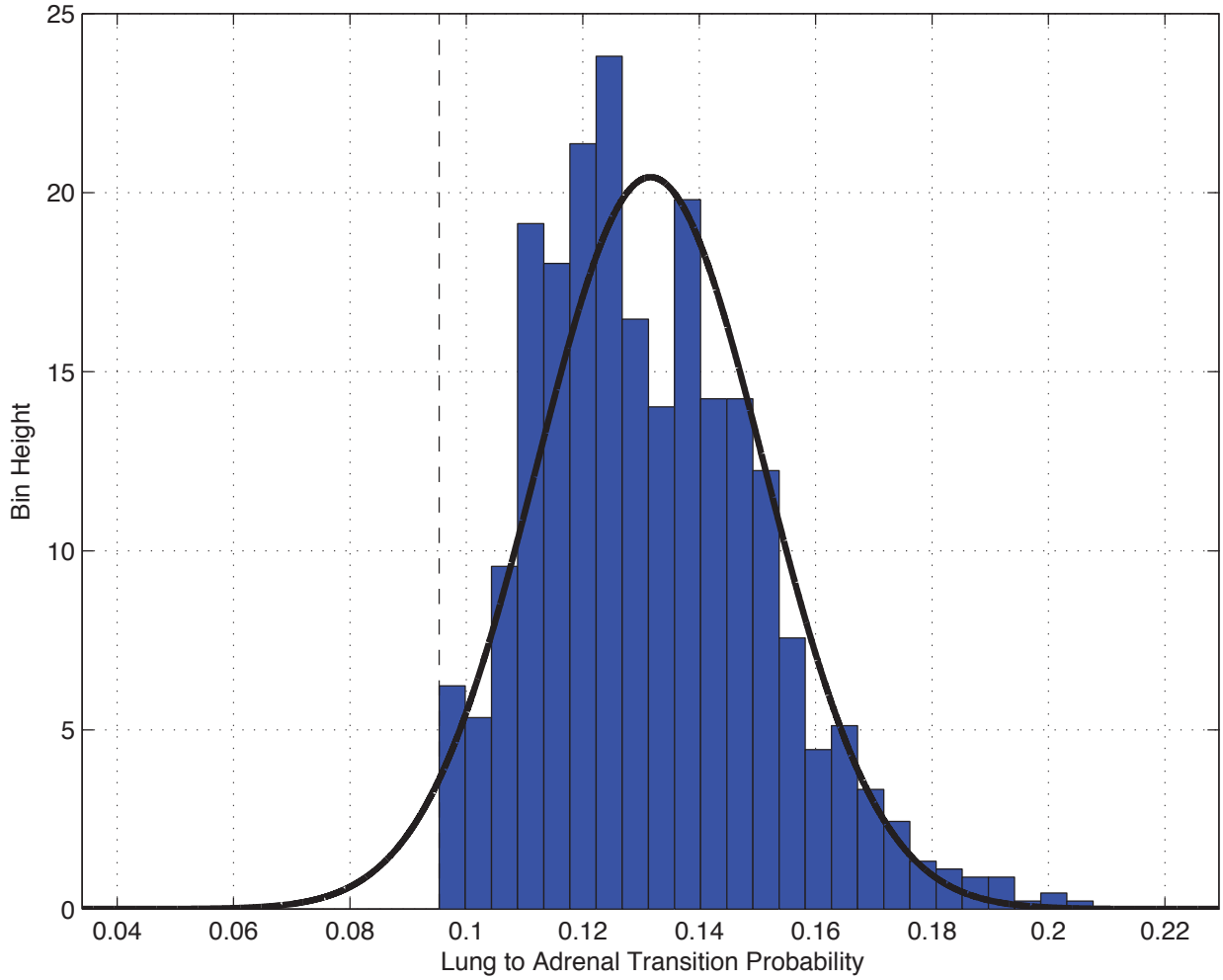


Figure 2.6: Lung → adrenal histogram
Histogram of edge values from lung to adrenal for 1000 trained $A_f$'s showing that edge values (transition probabilities) are best thought of as random variables which are (approximately) normally distributed. Dashed vertical line shows initial edge value associated with $A_0$. Normal distribution with sample mean (0.13165) and variance (0.01953) is shown as overlay.

(all are ensemble averaged) it is clear that the top 20 sites (listed above the cut-off line) have direct transition values higher than their most probable two-step transition, hence we call these 'first-order' sites. If the disease reaches one of these sites, the most likely path is directly from the lung after one-step. A random walker, leaving the lung site, after it chooses

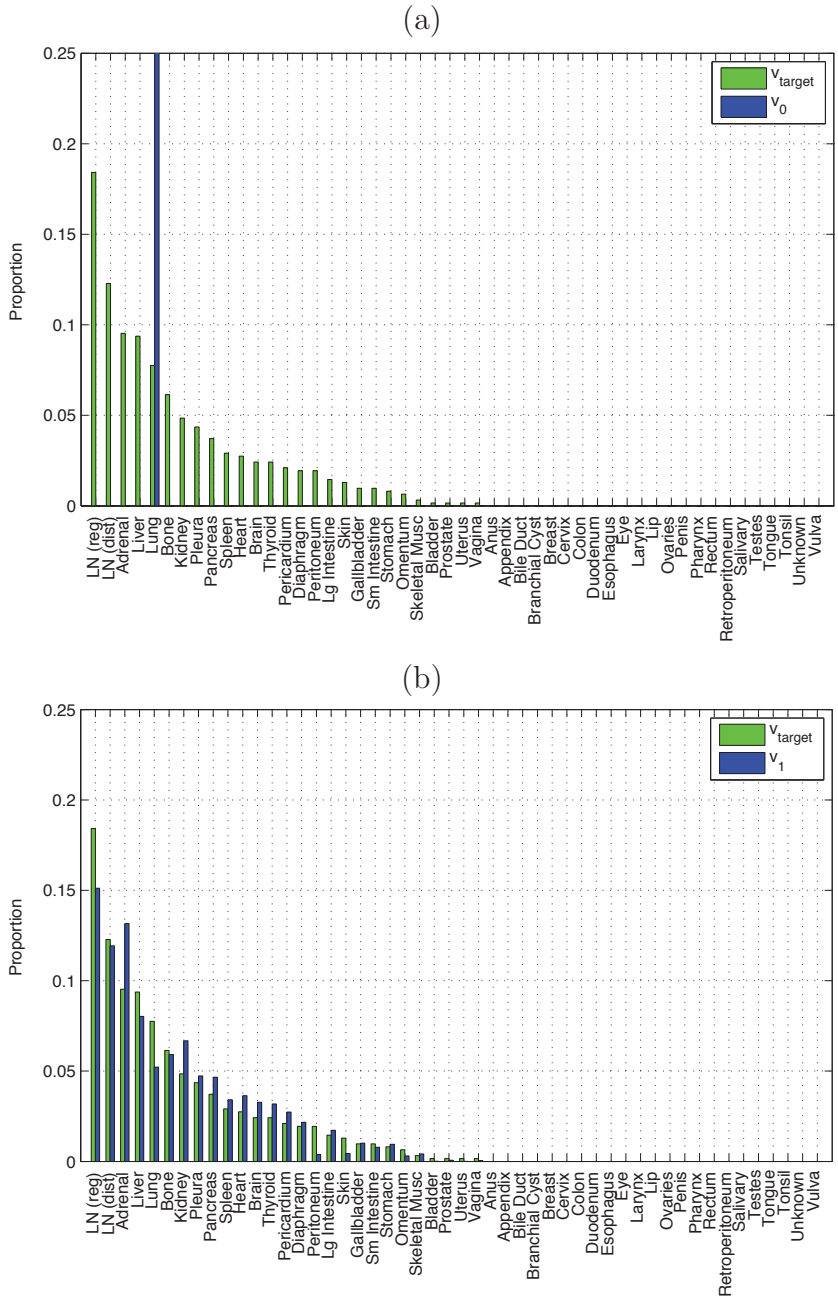


Figure 2.7: State vector progression for $k = 0$ & $k = 1$

Panel showing progression of state vector $\vec{v}_k$ for lung cancer primary using the ensemble averaged lung cancer matrix. Filled rectangles show the long-time metastatic distribution from the autopsy data in Figure 1.2(b), unfilled rectangles show the distribution at step $k$ using the Markov chain model. (a) $k = 0$; (b) $k = 1$.

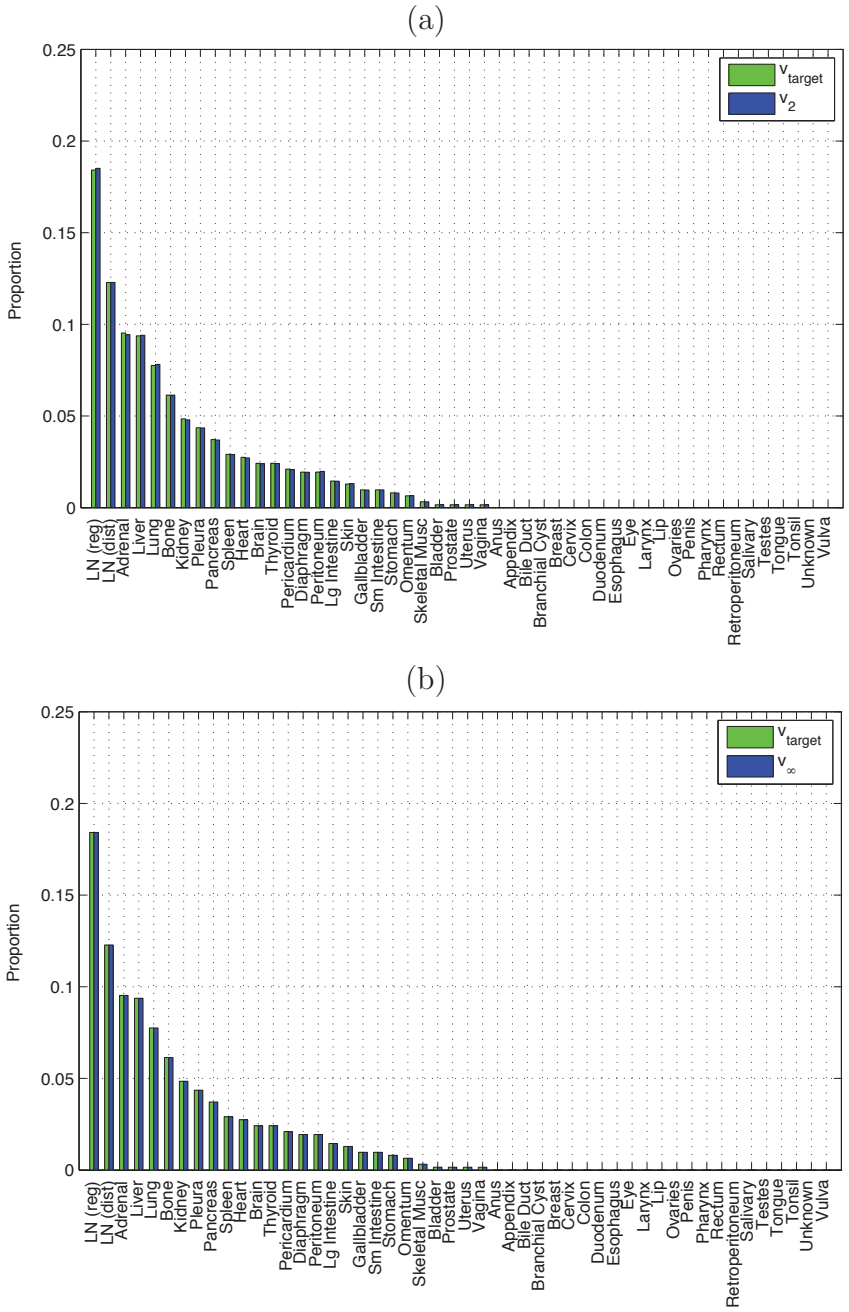


Figure 2.8: State vector progression for $k = 2$ & $k = \infty$

Panel showing progression of state vector $\vec{v}_k$ for lung cancer primary using the ensemble averaged lung cancer matrix. Filled rectangles show the long-time metastatic distribution from the autopsy data in Figure 1.2(b), unfilled rectangles show the distribution at step $k$ using the Markov chain model. (a) $k = 2$; (b) $k = \infty$.

one of the available outgoing edges with probability corresponding to the edge weighting, will first visit one of these first-order sites. The most heavily weighted edges, hence the most likely first site visits, will be regional lymph nodes and adrenal, accounting for roughly 28% of the first site visits. The next two most heavily weighted sites are distant lymph nodes and liver. These four sites account for roughly 50% of the first site visits of an ensemble of random walkers.

The remaining 7 sites (below the cut-off, starting from skin) have two-step transition path probabilities that are equal to or more probable than their direct one-step path from lung (taking into account standard deviations). We call these the 'second-order' sites. The interpretation of these sites is if there is a metastatic tumor at one of these sites, it is equally probable, or more probable that there is also a metastatic tumor at an intermediate site, most probably the regional lymph nodes or adrenal gland. Skin is the most significant second-order site, suggesting a possible pathway from a primary tumor in the lung to a metastatic tumor on the skin via the regional lymph nodes or adrenal gland (not shown, but almost as probable).

The classification of sites allows us to quantify possible disease progression paths (described in terms of 'random walkers') from lung to a given metastatic location. This is shown in Figure 2.9 where we focus on the multiple pathways by which cancer can spread from a primary lung tumor to the liver. We show in the figure the outgoing connection from lung to liver (with weight $0.08028 \pm 0.00946$), since liver is a first-order site. Roughly 92% of the random walkers, however, do not transition to liver on the first step, but go instead to a

Table 2.1: **One- and two-step transition probabilities**

| Target Sites | One-step transition prob (Avg) | Two-step transition probs |
|---|---|---|
| LN (reg) | 0.15115 ± 0.01821 | 0.02819 (LN (reg)) |
| Adrenal | 0.13165 ± 0.01953 | 0.01397 (LN (reg)) |
| LN (dist) | 0.11928 ± 0.00279 | 0.01860 (LN (reg)) |
| Liver | 0.08028 ± 0.00946 | 0.01440 (LN (reg)) |
| Kidney | 0.06677 ± 0.01231 | 0.00709 (LN (reg)) |
| Bone | 0.05914 ± 0.00196 | 0.00931 (LN (reg)) |
| Lung | 0.05223 ± 0.01504 | 0.01214 (LN (reg)) |
| Pleura | 0.04735 ± 0.00338 | 0.00657 (LN (reg)) |
| Pancreas | 0.04660 ± 0.00785 | 0.00549 (LN (reg)) |
| Heart | 0.03639 ± 0.00739 | 0.00407 (LN (reg)) |
| Spleen | 0.03415 ± 0.00454 | 0.00432 (LN (reg)) |
| Brain | 0.03274 ± 0.00728 | 0.00360 (LN (reg)) |
| Thyroid | 0.03180 ± 0.00628 | 0.00356 (LN (reg)) |
| Pericardium | 0.02733 ± 0.00557 | 0.00306 (LN (reg)) |
| Diaphragm | 0.02169 ± 0.00216 | 0.00289 (LN (reg)) |
| Lg Intestine | 0.01724 ± 0.00266 | 0.00219 (LN (reg)) |
| Gallbladder | 0.01015 ± 0.00048 | 0.00145 (LN (reg)) |
| Stomach | 0.00949 ± 0.00139 | 0.00119 (LN (reg)) |
| Sm Intestine | 0.00786 ± 0.00158 | 0.00149 (LN (reg)) |
| Skeletal Musc | 0.00413 ± 0.00093 | 0.00047 (LN (reg)) |
| Skin | 0.00439 ± 0.00443 | 0.00203 (LN (reg)) |
| Peritoneum | 0.00384 ± 0.00567 | 0.00308 (LN (reg)) |
| Omentum | 0.00305 ± 0.00223 | 0.00103 (LN (reg)) |
| Prostate | 0.00064 ± 0.00060 | 0.00025 (LN (reg)) |
| Vagina | 0.00052 ± 0.00059 | 0.00025 (LN (reg)) |
| Bladder | 0.00009 ± 0.00029 | 0.00023 (Adrenal) |
| Uterus | 0.00007 ± 0.00025 | 0.00022 (Adrenal) |

The 27 target sites listed in decreasing order of their edge weights (ensemble average values) from lung site. The 20 sites above the 'cut-off' are called 'first-order' sites. Their direct connections from the lung are strong enough so that they represent the most likely route to that site. The 7 sites listed below are called 'second-order' sites. Their connections from the lung are sufficiently weak that it is equally or more likely (taking into account standard deviations) to get to the site via some other first-order site (shown in parentheses).

different first-order site. Some of these will pass to the liver on the second step, as shown by the directed (solid) arrows. Still others pass to a second-order site, and then to the liver, as shown by the directed (dashed) arrows. In this way, all possible pathways to the liver from lung can be compared probabilistically and one can make quantitative predictions on which other sites might have metastases if a lung cancer patient develops a metastatic liver tumor.

### 2.3.5 Self-seeding sites

A recent focus in the literature has been on the possibility that tumors can 'self-seed' (see [39, 60]) since that process would help explain the exceptionally rapid ('Gompetzian' [59]) growth of certain primary tumors. In addition, these papers discuss the possibility, not yet proven experimentally, that self-seeding could potentially occur from a metastatic site back to itself, i.e. 'metastasis reseeding'. The focus on self-seeding of the primary tumor (CTCs that colonize their tumors of origin) demonstrated convincingly in mouse models [39] has led to the general concept that cancer progression, and hence progression pathways, may not be a strictly unidirectional process of progression from primary tumor to sequentially distant metastatic sites. It may well involve aspects that are more multidirectional in nature, such as tumor self-seeding, reseeding of the primary tumor from a metastatic tumor, or reseeding of a metastatic site from the metastatic tumor. Experimental evidence and the development of theoretical models that support this, is currently an active area of research. In our model, a site that is self-seeding is one in which a random walker leaving that site can return directly. The simplest way (but not the only way) to do this would be after one step, if the site

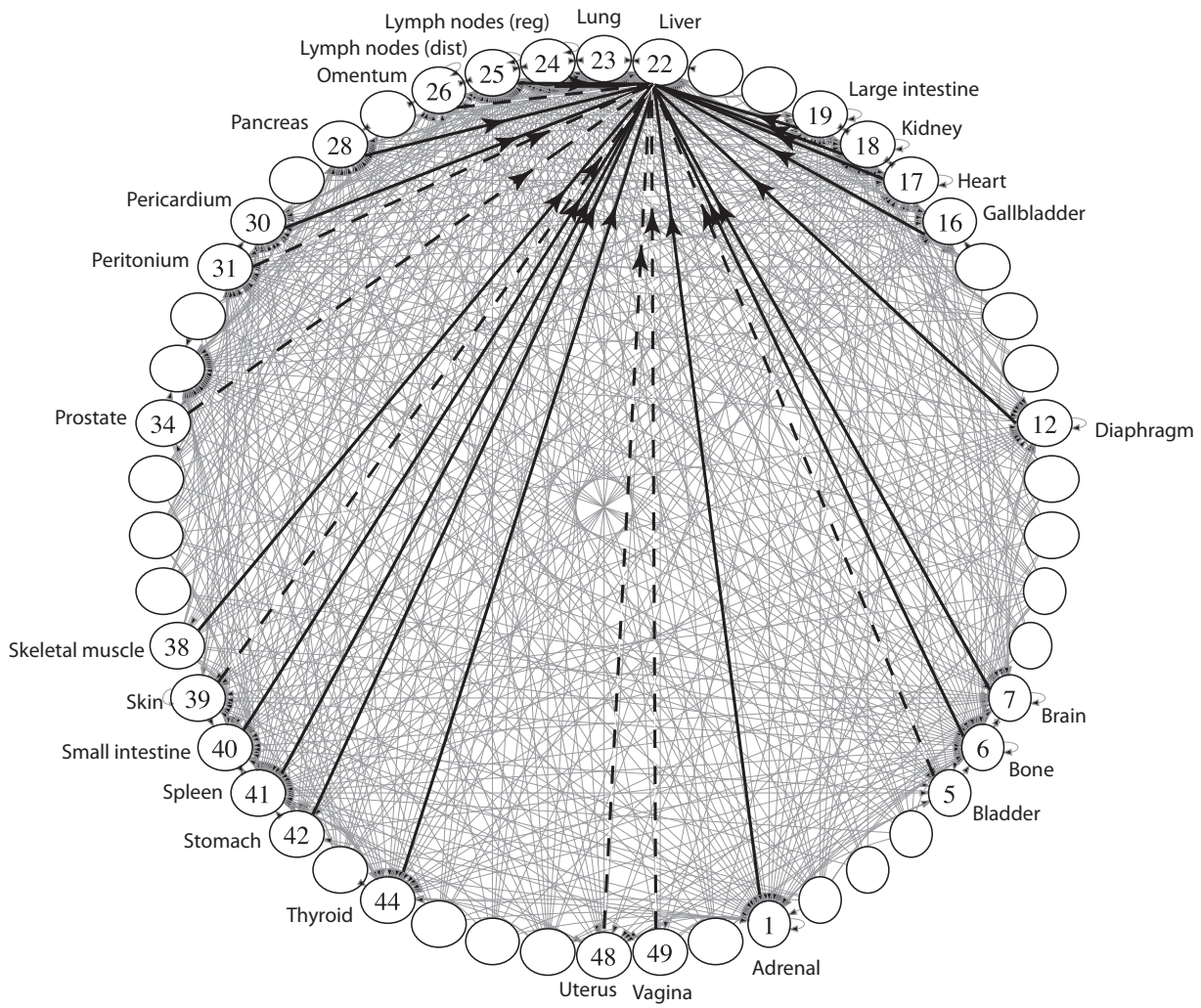


Figure 2.9: Pathways from lung to liver
Probabilistic decomposition of pathways from lung to liver. First transition probability is directly from lung to liver ($0.08028 \pm 0.00946$). Paths from the first-order sites to liver are shown as solid arrows. Paths from second-order sites to liver are shown as dashed arrows.

has an edge connecting back to itself. This would correspond to a non-zero probability in the diagonal entry of the transition matrix. We list in Table 2.2 the sites that have this property, along with the edge weighting, listed from strongest to weakest. For primary lung cancer, the most strongly weighted self-connecting edges are the lymph nodes (regional

and distant), liver, adrenal, bone, and lung. A more thorough analysis of this potentially important multidirectional mechanism of progression for each given type of primary cancer, along with the average time it takes to self-seed will be discussed later.

Table 2.2: **Self-edge weightings for each site**

| Target Sites | Self-edge weight (avg) |
|---|---|
| LN (reg) | 0.1865 ± 0.0152 |
| LN (dist) | 0.1231 ± 0.0028 |
| Liver | 0.0945 ± 0.0094 |
| Adrenal | 0.0929 ± 0.0212 |
| Bone | 0.0616 ± 0.0019 |
| Lung | 0.0522 ± 0.0150 |
| Kidney | 0.0470 ± 0.0143 |
| Pleura | 0.0434 ± 0.0049 |
| Pancreas | 0.0360 ± 0.0097 |
| Spleen | 0.0286 ± 0.0057 |
| Heart | 0.0262 ± 0.0088 |
| Thyroid | 0.0233 ± 0.0076 |
| Brain | 0.0230 ± 0.0092 |
| Peritoneum | 0.0211 ± 0.0122 |
| Pericardium | 0.0203 ± 0.0071 |
| Diaphragm | 0.0192 ± 0.0031 |
| Lg Intestine | 0.0141 ± 0.0033 |
| Skin | 0.0140 ± 0.0071 |
| Sm Intestine | 0.0098 ± 0.0019 |
| Gallbladder | 0.0097 ± 0.0007 |
| Stomach | 0.0081 ± 0.0019 |
| Omentum | 0.0068 ± 0.0030 |
| Skeletal Musc | 0.0032 ± 0.0013 |
| Bladder | 0.0020 ± 0.0025 |
| Uterus | 0.0020 ± 0.0025 |
| Vagina | 0.0017 ± 0.0012 |
| Prostate | 0.0017 ± 0.0009 |

27 target sites and their self edge weights (ensemble average) listed in decreasing order.

### 2.3.6 Mean first-passage times

An important quantity associated with our model is called 'mean first-passage time' to each of the sites - how many steps, on average, does it take for a random walker to pass from the lung site to each of the other sites. This gives us a model based timescale (not limited to take on discrete values) associated with disease progression, something a static autopsy data set cannot give us directly. It is important to keep in mind that these values are model based only, they do not arise from comparisons of disease time histories, something that could be done with a different data set that contains time progression information. To calculate these times, we follow a random walker starting at the lung position, progressing from site to site until all of the sites have been visited at least one time. Using this method for roughly 10,000 of these random walkers, we collect statistical information on the mean first-passage time to each of the sites, i.e. the average number of steps it takes to first arrive at each site. We show below in Table 2.3 the mean first-passage times from the lung site, which we obtain by Monte Carlo simulations using an ensemble of 10,000 realizations, where each realization is run long enough in time so that all sites identified by the lung cancer target vector are visited at least once. We emphasize that the mean first-passage times are distributed over a range of positive values quite distinct from the discrete values required in the underlying Markov process.

Despite the fact that these mean first-passage times are model-based (i.e. time passage information is not directly in the data set) they are interesting from several points of view. The normalized values, shown in the right column of the table, are obtained by dividing each

Table 2.3: **Mean first-passage times from lung**

| Target Sites | MFPT (un-normalized) | MFPT (normalized) |
|---|---|---|
| LN (reg) | 5.6414 ± 0.4919 | 1.0000 ± 0.0872 |
| LN (dist) | 8.3541 ± 0.8096 | 1.4809 ± 0.1435 |
| Adrenal | 10.0349 ± 1.0068 | 1.7788 ± 0.1785 |
| Liver | 10.6139 ± 1.0226 | 1.8814 ± 0.1813 |
| Lung | 13.0284 ± 1.1497 | 2.3094 ± 0.2038 |
| Bone | 16.0277 ± 1.4508 | 2.8411 ± 0.2572 |
| Kidney | 20.3944 ± 1.9664 | 3.6151 ± 0.3486 |
| Pleura | 22.9329 ± 2.4375 | 4.0651 ± 0.4321 |
| Pancreas | 26.4350 ± 2.6438 | 4.6859 ± 0.4686 |
| Spleen | 33.7009 ± 3.4925 | 5.9739 ± 0.6191 |
| Heart | 36.5513 ± 3.6359 | 6.4791 ± 0.6445 |
| Brain | 40.5540 ± 4.3179 | 7.1886 ± 0.7654 |
| Thyroid | 41.3240 ± 4.0700 | 7.3251 ± 0.7215 |
| Pericardium | 46.8599 ± 4.1645 | 8.3064 ± 0.7382 |
| Diaphragm | 51.3372 ± 5.6196 | 9.1001 ± 0.9961 |
| Peritoneum | 51.9555 ± 5.4518 | 9.2097 ± 0.9664 |
| Lg Intestine | 69.0501 ± 7.3192 | 12.2399 ± 1.2963 |
| Skin | 79.2006 ± 8.4505 | 14.0392 ± 1.4979 |
| Gallbladder | 104.9654 ± 10.0373 | 18.6063 ± 1.7792 |
| Sm Intestine | 105.8723 ± 9.9567 | 18.7670 ± 1.7649 |
| Stomach | 122.4070 ± 12.7034 | 21.6980 ± 2.2518 |
| Omentum | 155.6364 ± 15.8049 | 27.5883 ± 2.8016 |
| Skeletal Musc | 313.7172 ± 30.6400 | 55.6098 ± 5.4313 |
| Bladder | 620.7585 ± 63.7243 | 110.0362 ± 11.2958 |
| Prostate | 630.6260 ± 68.4618 | 111.7854 ± 12.1356 |
| Vagina | 630.8929 ± 64.6222 | 111.8327 ± 11.4550 |
| Uterus | 633.1578 ± 63.9966 | 112.2342 ± 11.3441 |

Mean first-passage times (un-normailzed and normalized) from lung to each target site, obtained by Monte Carlo simulation. Histogram plot is shown in Figure 2.10.

entry of the un-normalized column by the regional lymph node passage time time of 5.6414. This way, everything is measured with respect to the time associated with the progression from lung to regional lymph nodes, providing a relative predictive timescale for average progression. If a patient with a primary lung tumor progresses to a metastatic tumor in

the regional lymph nodes after one year, one might expect it to take roughly another 6 months to progress to the distant lymph nodes, or roughly 9 months to the adrenal gland. The interpretation is not that the disease will spread from lung to lymph nodes to liver to adrenal, etc. all in one individual patient (since the model is based on an ensemble data set), but that one, or perhaps several of these secondary sites will eventually produce metastatic tumors, and we have a predictive handle on the progression timescales. The mean first-passage time histogram is plotted in Figure 2.10 and gives a visual representation of the relative timescales to each of the sites. The sites seem to be grouped into approximately three clusters. In the first group, consisting of sites LN (reg) - Bone, there is an approximate linear increase in the mean first-passage times. The second grouping (Kidney - Peritoneum) also increases linearly, but on a slightly shifted line. The third grouping (Lg intestine - Uterus) increases (roughly) exponentially. Sites in this group, with very large mean first-passage times, like prostate or bladder, would be ones in which, if a metastatic tumor does appear, would indicate poor prognosis as other areas would have had a lot of time and 'probabilistic' opportunities to develop tumors as well.

Not shown in the table and figure are mean first-passage times from sites other than lung. But it is worth pointing out that we have calculated these times starting at all 50 sites, and the shortest mean first-passage time occurs from pleura to adrenal, with a un-normalized time of 1.02, or normalized value of 0.1811. This exceptionally short passage time indicates that if the lung tumor does progress to the pleura, one might expect a short time later for progression to occur to the adrenal gland. As mentioned earlier, this is another possible indication of

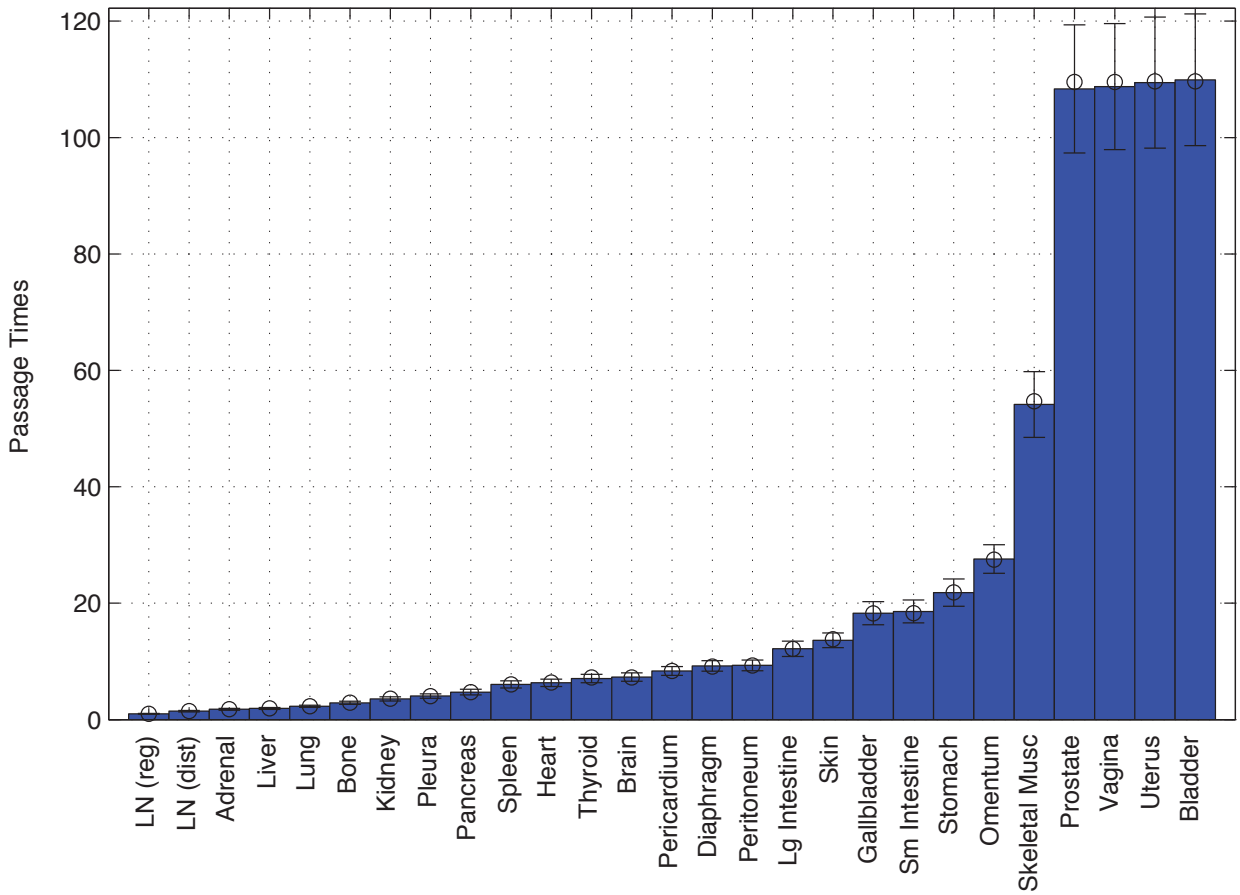


Figure 2.10: Mean first-passage time histogram
Mean first-passage time histogram for Monte Carlo computed random walks all starting from lung. Error bars show one standard deviation. Values are normalized so that LN (reg) has value 1, and all others are in these relative time units.

the potential importance of adrenal gland involvement in lung cancer progression. We are currently comparing our model based mean first-passage times with other data sets that contain the time-history of the disease in individual patients and ensembles.

## 2.4 Discussion

This chapter has set up the framework for the Markov model that will be used for detailed analysis, hypothesis testing, and comparative studies. After building a successful model, the network is analyzed from a mathematical standpoint and translated in such a way that gives clinicians useful information. Labeling and ordering the sites relate to the more and less important metastatic sites for the progression of the primary cancer, and timescales of progression illustrate timeframes to a metastasis. Even though the calculations were focused only on building a lung cancer model, it is important to know that these same techniques can be used for any primary cancer.

# Chapter 3

# Spreaders and sponges define metastasis in cancer

## 3.1 Introduction

This section focuses mainly on the key role that 'spreaders' and 'sponges' play in primary lung cancer. By rank-ordering the top two-step pathways emanating from the primary cancer, we can observe the main sites that dictate the spread of the disease and then classify these sites using their respective pathway probabilities. We then look at the timescales of progression given by the model (a reflection of the time taken for a primary tumor to metastasize) and compare it against the industry standard Kaplan-Meier curves. By assimilating new data into the model from [73], we further compare these timescales as well as the top two-step paths used to classify the 'spreaders' and 'sponges'.

## 3.2 Methods

### 3.2.1 Structure of the lung cancer multistep diagram

The 27 metastatic sites in the diagram shown in Figure 3.1 are organized in ring formation, with 20 sites surrounding lung on the inner ring and the remaining 7 sites organized on the outmost ring, each connected to a site from the inner ring. The sites listed on the inner ring are called 'first-order' sitesthey have direct edge connections from the lung, with edge probabilities decreasing from 12:00 clockwise around the ring. The most heavily weighted edge, hence the most likely first step of metastatic disease, is the transition from lung to regional lymph nodes [LN (reg)]. The least heavily weighted, hence least likely first step, is the transition from lung to skeletal muscle shown just to its left. The 7 sites making up the outermost ring are called 'second-order' sites, also organized with edge probabilities decreasing in clockwise order. These sites are classified as 'second-order' due to the fact that they have two-step probabilities via a first-order site that are equal or higher in probability than any direct one-step probability from the lung. In short, for disease to spread to a second-order site from lung, it most probably passes via a first-order site.

The general structure of the concentric diagram, with lung placed at the center, highlights the underlying classical uni-directional view of disease progression. However, the diagram also highlights the 3 key mechanisms of multidirectional progression: (i) self-seeding of the primary lung tumor shown in the diagram as a self-loop in the seventh position, with an edge weight of 5.2% and (ii) reseeding of the primary tumor from a first-order site, shown

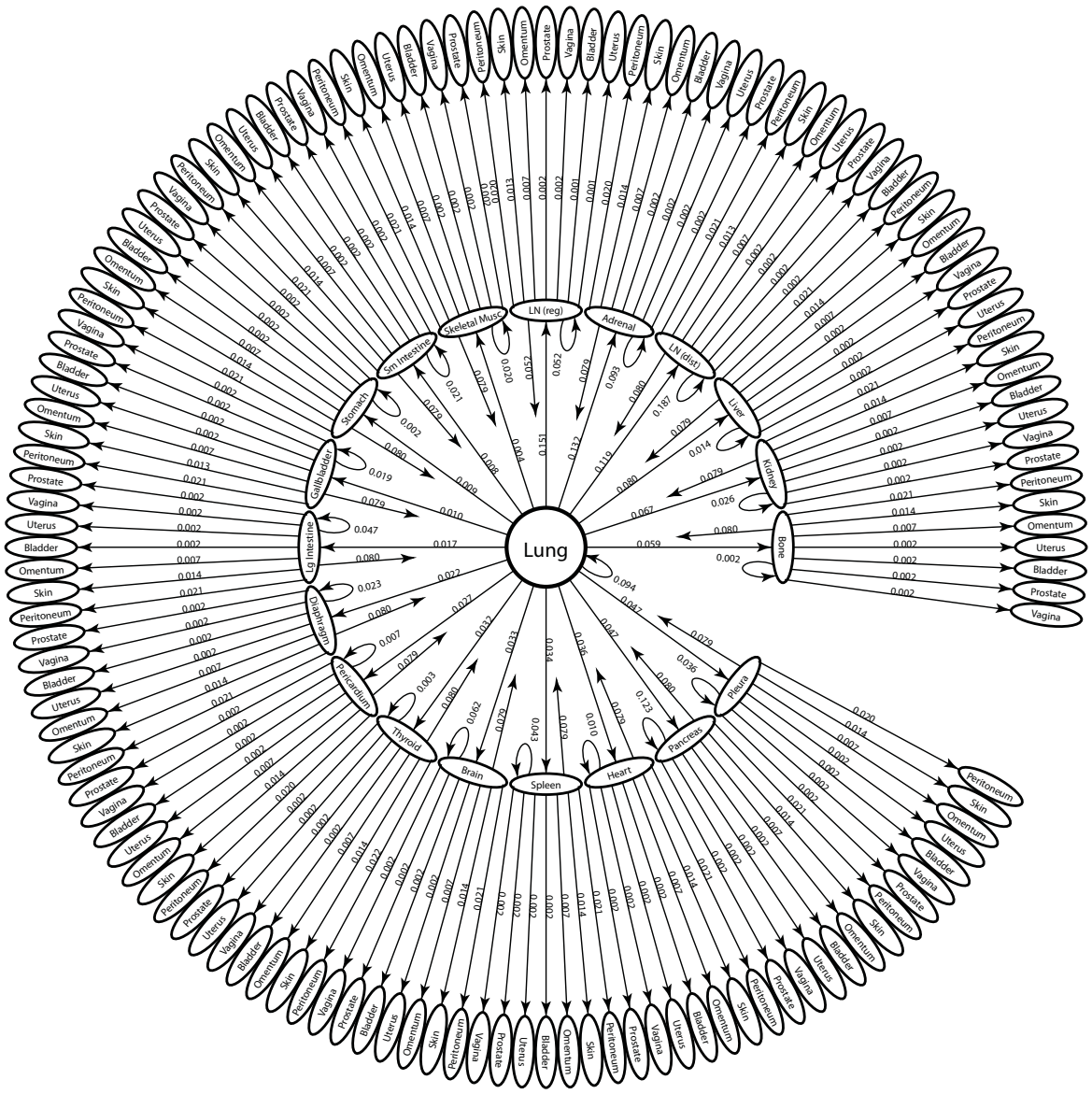


Figure 3.1: The one-step pathways of metastatic lung cancer

Ensemble averaged one-step pathway diagram. Primary lung tumor is at the center, next ring out are the 20 first-order sites showing their direct connection from the lung, with transition probabilities getting weaker in clockwise direction. Next ring out are the 7 second-order sites and their connections from the first-order sites. The 3 elements of multidirectional spread are highlighted in this diagram: (i) self-seeding of the primary tumor (self-loop back to center), (ii) reseeding of the primary tumor from a first-order site (arrows back to center), and (iii) reseeding of first-order sites (self-loops back to first-order site). Not shown in the diagram are the one-step paths from first-order site to another first-order site.

as arrows directed back to the center. Because we are using an ensemble average of 1,000 trained lung cancer matrices to produce this diagram, the reseeding edges are all roughly comparable in weight (8%), (iii) metastasis reseeding of first-order sites shown as a self-loop back to each metastatic site. The strongest metastasis re-seeders are lymph nodes (regional and distant), followed by liver, adrenal, bone, and kidney.

From this diagram, we can obtain the two-step pathway probabilities from the lung, by direct multiplication of the 2 edges making up any of the two-step paths starting from lung. The 729 distinct two-step paths from the lung, the top ones of which are shown in Figure 3.2, produce the statistical distribution $\vec{v}_2$ produced by the Markov chain model. We calculate all of these and rank them in decreasing order in the next subsections. By comparing the probability distributions $\vec{v}_2$ and $\vec{v}_\infty$ (shown in Figure 2.8a), we can see that after 2 steps, the distribution has nearly converged to the steady-state, so we expect our rankings of two-step pathways not to change much if we compare them to the top three-step and higher step pathways.

Figure 3.3 shows a (ensemble) convergence and non-convergence plot associated with our search algorithm to calculate the Markov transition matrix based on the baseline dataset [19]. What is significant is the non-convergence of our algorithm when we constrain our searches to not allow for any multidirectional edges. In other words, when we forced our algorithm to not allow edges directly back to a site (no self-metastases nor primary reseeding), either separately or together, the algorithm would not converge to a solution. In contrast, the

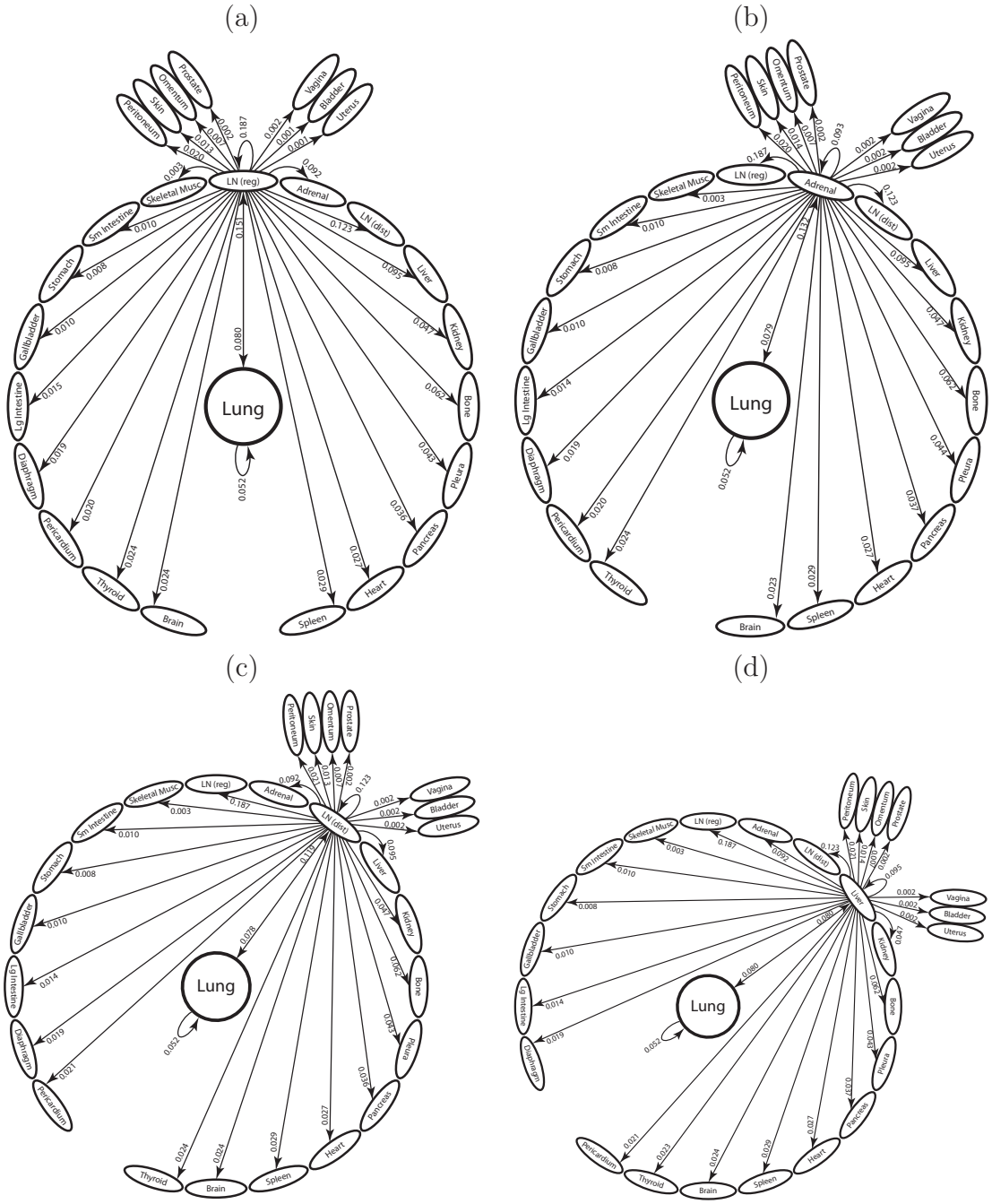


Figure 3.2: The two-step pathways through top 8 first-order sites
Diagram of all 28 two-step pathways from lung to a tertiary site. a. lung through regional lymph nodes. b. lung through adrenal gland. c. lung through distant lymph nodes. d. lung through liver. (*Continued on the following page.*)

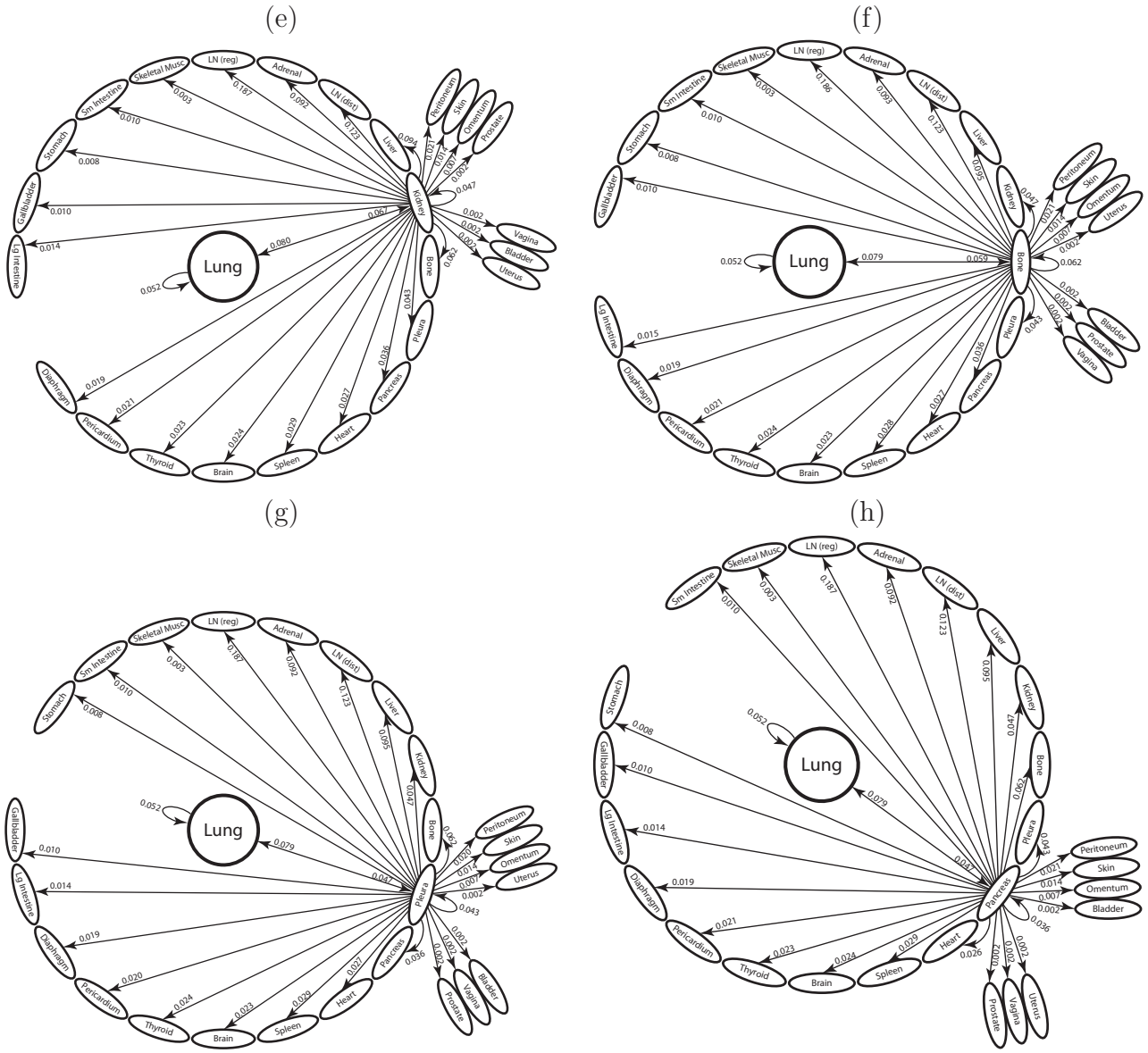


Figure 3.2 (*Continued.*)

e. lung through kidney. f. lung through bone. g. lung through pleura. h. lung through pancreas.

algorithm, in general, converged quickly to a solution when all connections were allowed and produces a transition matrix with many multidirectional connections from site to site.

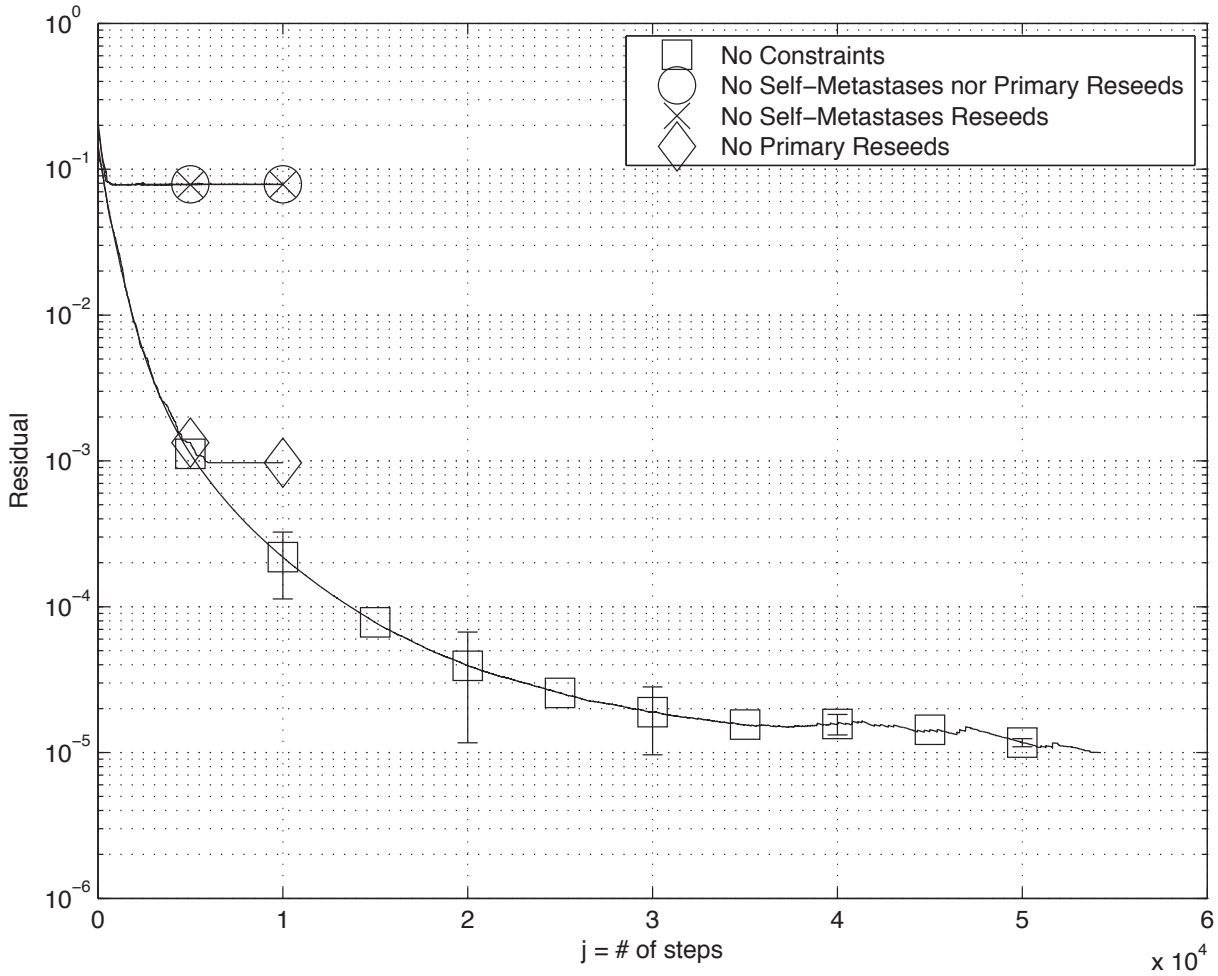


Figure 3.3: Convergence plot for the lung cancer matrix
Ensemble averaged (1000 trained matrices) convergence plot associated with algorithm to compute the lung cancer transition matrix. Curve marked with squares shows the convergence (ensemble averaged convergence plots) with no constraints to the lung cancer transition matrix used in this study. Curve marked with circles shows non-convergence when the search is constrained so that metastasis re-seeding and primary re-seeding are not allowed. Curve marked with X's shows non-convergence when the search is constrained so that only metastasis re-seeding is not allowed. Curve marked with diamonds shows non-convergence when the search is constrained so that only primary re-seeding is not allowed.

### 3.2.2 The autopsy datasets

The data in [19] compiles the metastatic tumor distributions in a population of 3,827 deceased cancer patients, none of whom received chemotherapy or radiation, hence the model can be

said to be based on the natural progression of the disease, although mastectomy for many breast cancer primaries was most likely conducted at that time. In addition, brain metastases are likely underrepresented by this dataset as brain autopsies probably were not universally conducted at that time. The autopsies were conducted between 1914-1943 in 5 separate affiliated centers, with an ensemble distribution of 41 primary tumor types and 30 distinct metastatic locations. The total number of distinct primary and metastatic tumor locations is 50, which sets the size of our square Markov transition matrix ($50 \times 50$) as well as the number of entries in the Markov state vector $\vec{v}_k$. The data offer no direct information on the time history of the disease, either for individual patients comprising the ensemble or in ensemble format. The data we use, therefore, only contain information on the 'long-time' distribution of metastatic tumors, where long-time is associated with end of life, a timescale that varies significantly from patient to patient. The model does, however, allow us to infer time histories from autopsy data based on the logic that if more metastatic tumors show up in a population of patients at a specific site, then on average, they would develop earlier in the progression history. Although this association is not perfect, if does allow us to extract meaningful temporal inferences from our Markov chain model. Details of how we infer the correct ensemble Markov transition matrix are described in reference [54].

We use the dataset in 2 distinct ways to construct our model. First, we associate the distribution of metastatic tumors (after appropriate normalization) for primary patients with lung cancer with the steady-state (long-time) probability distribution of our Markov chain

[58]. From this, we compute the 'transition matrix' for our Markov chain (ensemble averaged) that produces this steady-state. As the problem is mathematically underdetermined, the calculation procedure requires an initial 'candidate' transition matrix obtained from the autopsy data and discussed in [54], which is then systematically iterated until a numerical convergence criterion is satisfied. Interestingly, we also show that when our search algorithm is con- strained so as to not allow any multidirectional edges in the directed graph associated with the transition matrix, no self- consistent model can be produced (i.e., the search algorithm does not converge). Then, we update our baseline model with the more targeted dataset described [73] of 137 patients with adenocarcinoma of the lung (stage I and II), all treated with complete lung resection, and show how the baseline model is able to adapt to this assimilated dataset.

## 3.3 Results

### 3.3.1 Cancer metastasis as a stochastic multistep process

The ensemble averaged lung cancer transition matrix associated with the Markov chain model (see Figure 3.1) depicts the complete metastatic pathway diagram [54]. Each of the 2,500 entries, $a_{ij}$, of the $50 \times 50$ transition matrix determines the probability of the disease (modeled as a random walker over the network) spreading from site '$i$' to site '$j$' in an effectively multistep process before the statistical tumor distribution of the autopsy dataset is filled out. The diagram rank orders (in decreasing clockwise order) all of the possible

pathways emanating from the lung. One-step paths are defined by the edges leading directly out from the lungthe sum of these outgoing edges must be one. The single most likely one-step path of disease progression from the lung is to the regional lymph nodes, shown at the top of the diagram, with a probability of 15.1%, followed by the lung to adrenal gland path, with probability of 13.2%. On the diagram ordering the first steps out of the lung, we also show the 'self-seeding' step directly back to the lung, represented by the edge from lung looping back to itself, with edge probability 5.2%.

Two-step paths are made up of an edge from the lung to another site (or back to itself), followed by the edge from that site to a second site. There are 729 two-step paths emanating from the lung. The probability of taking a particular two-step path from the lung is obtained by multiplying the weights of the 2 edges making up the path. The sum of all of these two-step path probabilities must be 1, and so on for three-step paths, four-step paths, etc. We focus on quantifying all of the two-step paths in this article, because as shown in Figure 2.8a, after 2 iterations of the Markov chain ($k = 2$), the state vector has nearly converged to the steady-state target vector for metastatic tumors making metastatic progression for lung cancer effectively a two-step process. In Figure 3.2, we show all of the two-step paths emanating from the lung passing through each of the 8 most probable metastatic sites. To obtain the probability of cancer progression on 1 of these two-step paths, one multiplies the products of the 2 edges making up the two-step path.

### 3.3.2 Rank-ordering the two-step metastatic pathways toward the final state of the disease

We list the top multidirectional two-step pathways obtained from our model in Table 3.1. The first entries of Table 3.1 list the top 10 reseeding pathways back to the lung from a first-order site, along with the running cumulative values. We highlight from this list several points. First, lymph nodes, adrenal gland, and liver are the most important intermediate sites that reseed back to the lung. Their cumulative probability value (3.8%) accounts for more than half of the total cumulative value from the entire list (6.2%). This total cumulative value is slightly greater than, but roughly comparable in size to the lung to lung reseeding path value of 5.2%, indicating that cells that reseed to the lung land therewith roughly equal probabilities of having arrived via an intermediate site (see Table 3.1) versus directly from the lung. The second half of Table 3.1 lists the top 10 two-step reseeding pathways back to a metastatic site, a mechanism we call 'metastasis reseeding.' From this table, we can see that for lung cancer, lymph nodes and adrenal gland are the most active metastasis re-seeders, followed by liver, bone, and kidney.

### 3.3.3 Metastatic sites as spreaders or sponges

A careful analysis of the top 30 two-step pathways allows us to compute the key probabilistic quantity of interest associated with each two-step path which characterizes each site as a sponge or a spreader. The quantity is the ratio of probability out ($P_{out}$) over probability in ($P_{in}$) to each of the sites. If $P_{out} > P_{in}$, the site is a spreader, whereas if $P_{in} > P_{out}$, we

Table 3.1: **Top two-step pathway probabilities**

| Top reseeding pathways back to lung | | | | | Transition probability | Cumulative values |
|---|---|---|---|---|---|---|
| Lung | → | LN (reg) | → | Lung | 0.01214 | |
| Lung | → | Adrenal | → | Lung | 0.01042 | 0.02256 |
| Lung | → | LN (dist) | → | Lung | 0.00952 | 0.03208 |
| Lung | → | Liver | → | Lung | 0.00645 | 0.03853 |
| Lung | → | Kidney | → | Lung | 0.00533 | 0.04386 |
| Lung | → | Bone | → | Lung | 0.00467 | 0.04853 |
| Lung | → | Pleura | → | Lung | 0.00375 | 0.05228 |
| Lung | → | Pancreas | → | Lung | 0.00367 | 0.05595 |
| Lung | → | Heart | → | Lung | 0.00288 | 0.05883 |
| Lung | → | Lung | → | Lung | 0.00273 | 0.06156 |
| Top metastasis reseeders | | | | | Transition probability | Cumulative values |
| Lung | → | LN (reg) | → | LN (reg) | 0.02819 | |
| Lung | → | LN (dist) | → | LN (dist) | 0.01468 | 0.04287 |
| Lung | → | Adrenal | → | Adrenal | 0.01223 | 0.05510 |
| Lung | → | Liver | → | Liver | 0.00758 | 0.06268 |
| Lung | → | Bone | → | Bone | 0.00364 | 0.06632 |
| Lung | → | Kidney | → | Kidney | 0.00314 | 0.06946 |
| Lung | → | Pleura | → | Pleura | 0.00206 | 0.07152 |
| Lung | → | Pancreas | → | Pancreas | 0.00168 | 0.07320 |
| Lung | → | Spleen | → | Spleen | 0.00098 | 0.07418 |
| Lung | → | Heart | → | Heart | 0.00095 | 0.07513 |

Top two-step reseeding pathways back to lung: Primary → First-order site → Primary. Top reseeding pathways back to metastatic site: Primary → First-order site → Back to first-order site. Cumulative values (obtained by adding the previous transition probabilities) are listed in third column.

characterize it as a sponge. The ratio ($P_{out}/P_{in}$) of their exiting and incoming probabilities, in the case of a spreader, gives us what we call the amplification factor, as it is larger than one, whereas in the case of a sponge, we call the ratio the absorption factor, as it is less than 1. Using these quantities, the top 2 spreaders are the adrenal gland and kidney, with amplification factors of 1.91 (adrenal gland) and 2.86 (kidney). The total number of two-step pathways into and out of the adrenal gland is 10, whereas the total into and out of kidney is

only 3. For these reasons, we identify the adrenal gland as the key distant anatomic spreader of primary lung cancer.

The sponges associated with primary lung cancer are the regional lymph nodes, liver, and bone. Their respective absorption factors are 0.74 (regional lymph nodes), 0.87 (liver), and 0.75 (bone). The total number of two-step pathways into and out of the regional lymph nodes is 16, compared with 8 into and out of the liver, and 5 into and out of bone. For these reasons, we identify the regional lymph nodes as the key anatomical sponge associated with primary lung cancer, followed by both bone and liver.

### 3.3.4 The spatial pathways of lung cancer

To compare the relative importance of two-step unidirectional pathways versus two-step multidirectional pathways, we list the top 30 two-step pathways in decreasing order in Table 3.2. The top metastatic pathway (of any type) is the lung $\rightarrow$ LN (reg) $\rightarrow$ LN (reg) metastasis reseeding pathway, whereas the top unidirectional pathway is the lung $\rightarrow$ adrenal $\rightarrow$ LN (reg) path. Looking at all of the multidirectional pathways, it is clear that the lymph nodes and adrenal gland are the key metastatic sites responsible for multidirectional spread, whereas lymph nodes, adrenal gland, and liver are important sites responsible for unidirectional spread. In general terms, lymph nodes, adrenal gland, and liver feature very prominently as intermediate metastatic sites in many of the two-step pathways.

The information can then be combined into a reduced 2-step diagram for progression, shown in Figure 3.4. The diagram shows the centrality of lymph nodes and adrenal gland

Table 3.2: **Comparative table of top two-step metastatic pathways of all types from Lung**

| | Top 30 Untreated/Baseline | | | Top 30 Stage I | | | Top 30 Stage II | | | Untreated | Stage I | Stage II |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *LN (reg) | → | LN (reg) | *LN (reg) | → | LN (reg) | *LN (reg) | → | LN (reg) | 0.02819 | 0.02766 | 0.02764 |
| 2 | Adrenal | → | LN (reg) | Adrenal | → | LN (reg) | Adrenal | → | LN (reg) | 0.02461 | 0.02320 | 0.02408 |
| 3 | LN (dist) | → | LN (reg) | LN (dist) | → | LN (reg) | LN (dist) | → | LN (reg) | 0.02234 | 0.01937 | 0.02057 |
| 4 | LN (reg) | → | LN (dist) | LN (reg) | → | LN (dist) | LN (reg) | → | LN (dist) | 0.01860 | 0.01613 | 0.01712 |
| 5 | Adrenal | → | LN (dist) | Liver | → | LN (reg) | Liver | → | LN (reg) | 0.01620 | 0.01550 | 0.01566 |
| 6 | Liver | → | LN (reg) | LN (reg) | → | Liver | Adrenal | → | LN (dist) | 0.01501 | 0.01476 | 0.01491 |
| 7 | *LN (dist) | → | LN (dist) | *LN (reg) | → | Lung | LN (reg) | → | Liver | 0.01468 | 0.01461 | 0.01488 |
| 8 | LN (reg) | → | Liver | Adrenal | → | LN (dist) | *LN (reg) | → | Lung | 0.01440 | 0.01349 | 0.01385 |
| 9 | LN (reg) | → | Adrenal | LN (reg) | → | Adrenal | LN (reg) | → | Adrenal | 0.01397 | 0.01283 | 0.01355 |
| 10 | Adrenal | → | Liver | Kidney | → | LN (reg) | Adrenal | → | Liver | 0.01253 | 0.01282 | 0.01292 |
| 11 | Kidney | → | LN (reg) | Adrenal | → | Liver | *LN (dist) | → | LN (dist) | 0.01245 | 0.01232 | 0.01271 |
| 12 | *Adrenal | → | Adrenal | *Adrenal | → | Lung | Kidney | → | LN (reg) | 0.01223 | 0.01209 | 0.01241 |
| 13 | *LN (reg) | → | Lung | *LN (dist) | → | LN (dist) | *Adrenal | → | Adrenal | 0.01214 | 0.01125 | 0.01185 |
| 14 | LN (dist) | → | Liver | Bone | → | LN (reg) | *Adrenal | → | Lung | 0.01130 | 0.01115 | 0.01184 |
| 15 | LN (dist) | → | Adrenal | *Adrenal | → | Adrenal | Bone | → | LN (reg) | 0.01101 | 0.01083 | 0.01109 |
| 16 | Bone | → | LN (reg) | *Lung | → | LN (reg) | LN (dist) | → | Liver | 0.01100 | 0.01042 | 0.01097 |
| 17 | *Adrenal | → | Lung | LN (dist) | → | Liver | *LN (dist) | → | Lung | 0.01042 | 0.01023 | 0.01019 |
| 18 | Liver | → | LN (dist) | *LN (dist) | → | Lung | LN (dist) | → | Adrenal | 0.00988 | 0.01013 | 0.01003 |
| 19 | *LN (dist) | → | Lung | LN (reg) | → | Bone | Liver | → | LN (dist) | 0.00952 | 0.00940 | 0.00970 |
| 20 | LN (reg) | → | Bone | Brain | → | LN (reg) | *Lung | → | LN (reg) | 0.00931 | 0.00924 | 0.00965 |
| 21 | Pleura | → | LN (reg) | Liver | → | LN (dist) | Pleura | → | LN (reg) | 0.00886 | 0.00904 | 0.00944 |
| 22 | Pancreas | → | LN (reg) | LN (dist) | → | Adrenal | LN (reg) | → | Bone | 0.00873 | 0.00894 | 0.00937 |
| 23 | Kidney | → | LN (dist) | Pleura | → | LN (reg) | *Lung | → | Adrenal | 0.00820 | 0.00886 | 0.00839 |
| 24 | Adrenal | → | Bone | *Lung | → | Adrenal | *Liver | → | Liver | 0.00811 | 0.00872 | 0.00835 |
| 25 | *Lung | → | LN (reg) | *Liver | → | Liver | Adrenal | → | Bone | 0.00789 | 0.00819 | 0.00815 |
| 26 | *Liver | → | Liver | *Liver | → | Lung | Pancreas | → | LN (reg) | 0.00758 | 0.00817 | 0.00814 |
| 27 | Liver | → | Adrenal | Adrenal | → | Bone | *Liver | → | Lung | 0.00735 | 0.00787 | 0.00781 |
| 28 | LN (dist) | → | Bone | Pancreas | → | LN (reg) | Kidney | → | LN (dist) | 0.00734 | 0.00775 | 0.00769 |
| 29 | Bone | → | LN (dist) | Kidney | → | LN (dist) | Liver | → | Adrenal | 0.00728 | 0.00748 | 0.00759 |
| 30 | LN (reg) | → | Kidney | LN (reg) | → | Kidney | *Lung | → | LN (dist) | 0.00709 | 0.00740 | 0.00715 |

Comparative table of top two-step metastatic pathways of all types from Lung. * paths are multidirectional. See Figure 3.1 for corresponding diagram. Since each path starts from the lung, we show only the 2nd and 3rd site in the two-step pathway. First column lists the two steps out from lung according to the baseline untreated dataset [19]. Columns 2 and 3 list the two steps out from lung according to the assimilated model which incorporates dataset [73]. Columns 4-6 list the two-step probabilities of the corresponding pathways.

as key first met- astatic sites, with many incoming and outgoing edges. The figure also captures all of the information about the spreader or sponge character of each site, with red indicating the color of the key spreaders (adrenal gland, kidney) and blue indicating the color of sponges (lung, regional lymph nodes, liver, bone). Amplification and absorption factors are shown in each of the ovals.

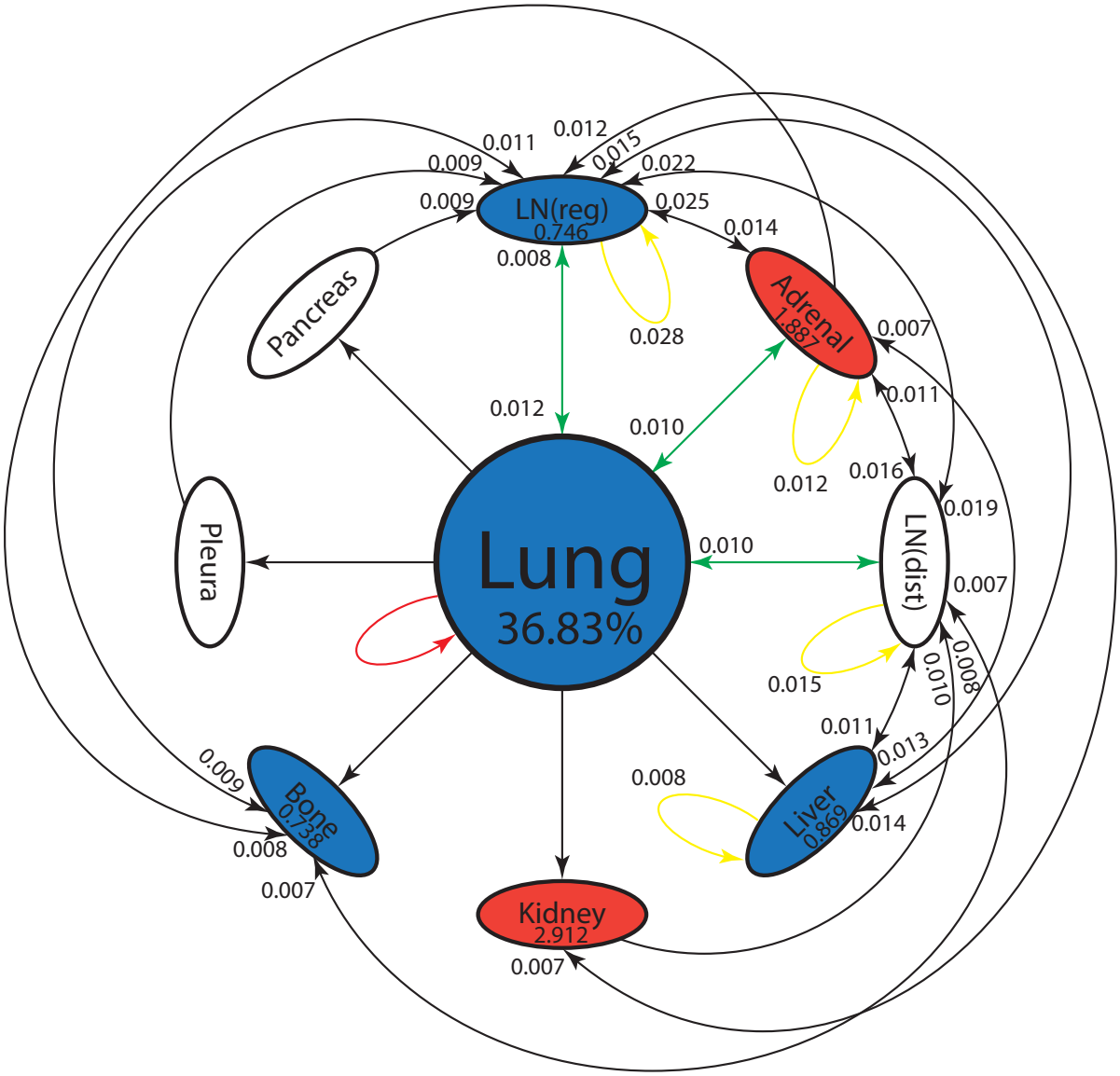


Figure 3.4: Reduced pathway diagram showing top 30 two-step paths

Top 30 two-step pathways emanating from lung (representing 36.83% of the total pathway probabilities), obtained by multiplying the edges of the one-step edges comprising each two-step path. Edges without numbers are one-step paths emanating from lung. All other numbered edges mark the second edge in a two-step path, with numbers indicating the two-step probabilities. Colors indicate classification of each node as a 'spreader' (red) or 'sponge' (blue). Spreader amplification factor and sponge absorption factor are listed in each oval. Edge colors indicate primary self-seeding (red), primary reseeding (green), and metastasis reseeding (yellow). See text for more detailed descriptions.

### 3.3.5 Timescales of progression: enhancing the Kaplan-Meier approach

Our model gives a useful measure of metastatic progression timescale, called first-passage time from lung to any given site, defined as the number of edges a 'random walker' leaving the lung must traverse to first arrive at that site. Monte Carlo simulations of random walk paths from the lung are conducted computationally to obtain mean first-passage times (averages over 10,000 runs) to every other site in the model. The mean first-passage times (MFPTs) act as a proxy timescale (model-based) for metastatic progression. It is a model-based relative measure of the time that it takes for a primary tumor to metastasize to a secondary site, or, roughly speaking, a model-based measure of the timescale associated with successful extravasation and colonization [23]. Timescales associated with metastatic disease are typically quantified by so-called Kaplan-Meier survival curves [43, 57], which follow a cohort of patients from presentation until death, plotting the survival percentage associated with the cohort (example shown in Figure 3.5). Alternative methods have been proposed, but by and large, tracking survival of a cohort of patients remains the industry-standard way of tracking progression. There is very little in the literature that tracks the timescale of progression from metastatic site to metastatic site [6, 21, 32, 36, 81].

Mean first-passage times from lung to each of the other sites are shown in Figure 3.6. The sites are ordered from shortest to longest mean first-passage time from lung. In dark, we show the baseline (untreated patients) model using the dataset [19]. The dashed-dot line is a linear curve fit to the first 9 sites, showing a clear linear increasing regimen (roughly the top

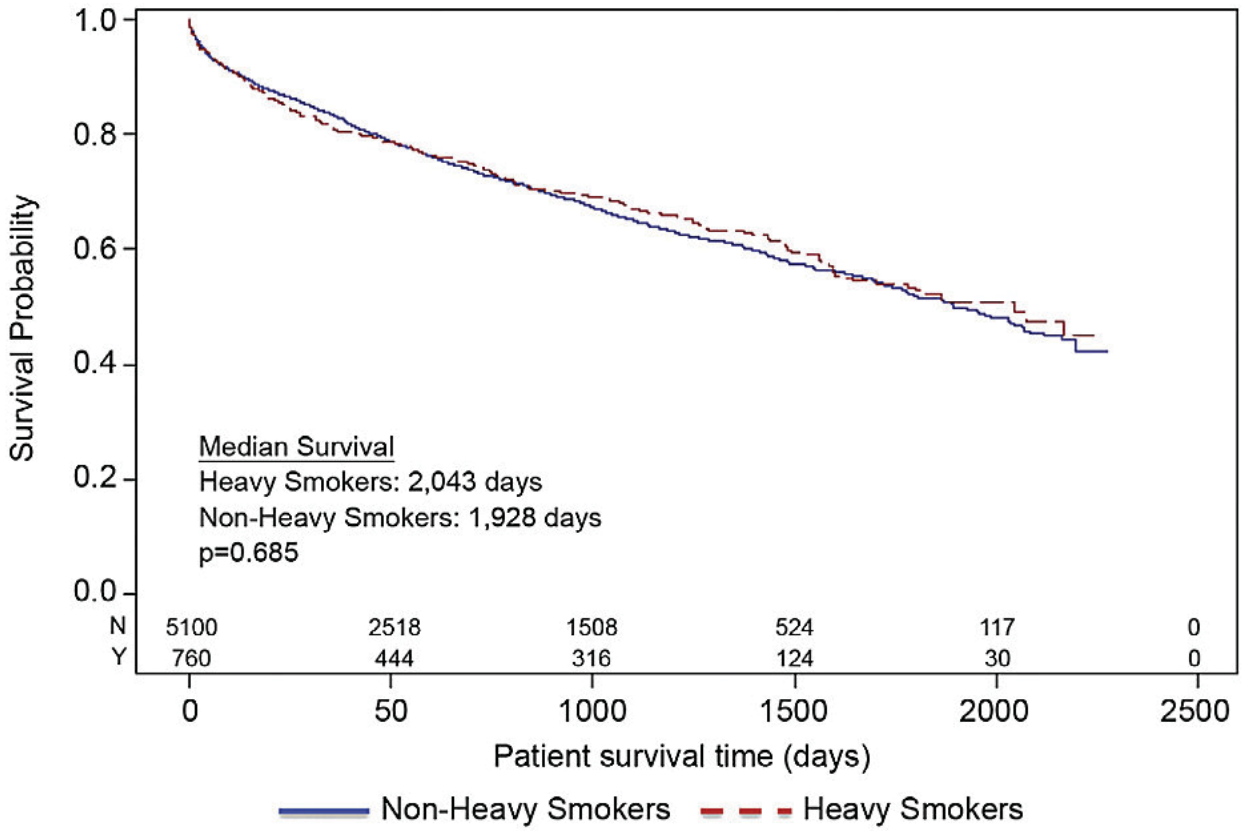


Figure 3.5: Kaplan-Meier curve of lung cancer victims
Example of an industry standard Kaplan-Meier survival curve. Curve shows the survival rate of lung cancer patients over a period of $\sim 2000$ days. Solid blue curve corresponds to patients that were considered non-heavy smokers. Dashed red curve corresponds to patients who were considered heavy smokers.

16 sites), followed by a group of sites where mean first-passage times increase nonlinearly. The first 9 sites used in the reduced model set the basic linear timescales of progression for the high probability metastatic locations. Times increase following the general linear formula $MFPT = a \cdot t + b$, where $a = 2.56, b = 2.07$ for the baseline (untreated) model, where '$a$' is the slope and '$b$' is the y-intercept. In this formula, larger slopes indicate longer overall mean first-passage times from lung to metastatic sites. Spread to regional lymph nodes is

fastest (with a normalized value of 1), followed by normalized times to distant lymph nodes (1.47), adrenal (1.72), and liver (1.75). One should interpret these timescales to indicate that it takes roughly 75% longer for cancer to metastasize to adrenal gland and liver than to regional lymph nodes. Self-seeding back to lung has a normalized mean first-passage time of 2.30, which is faster than to most of the first-order sites, but over twice the time as the lung to regional lymph node timescale.

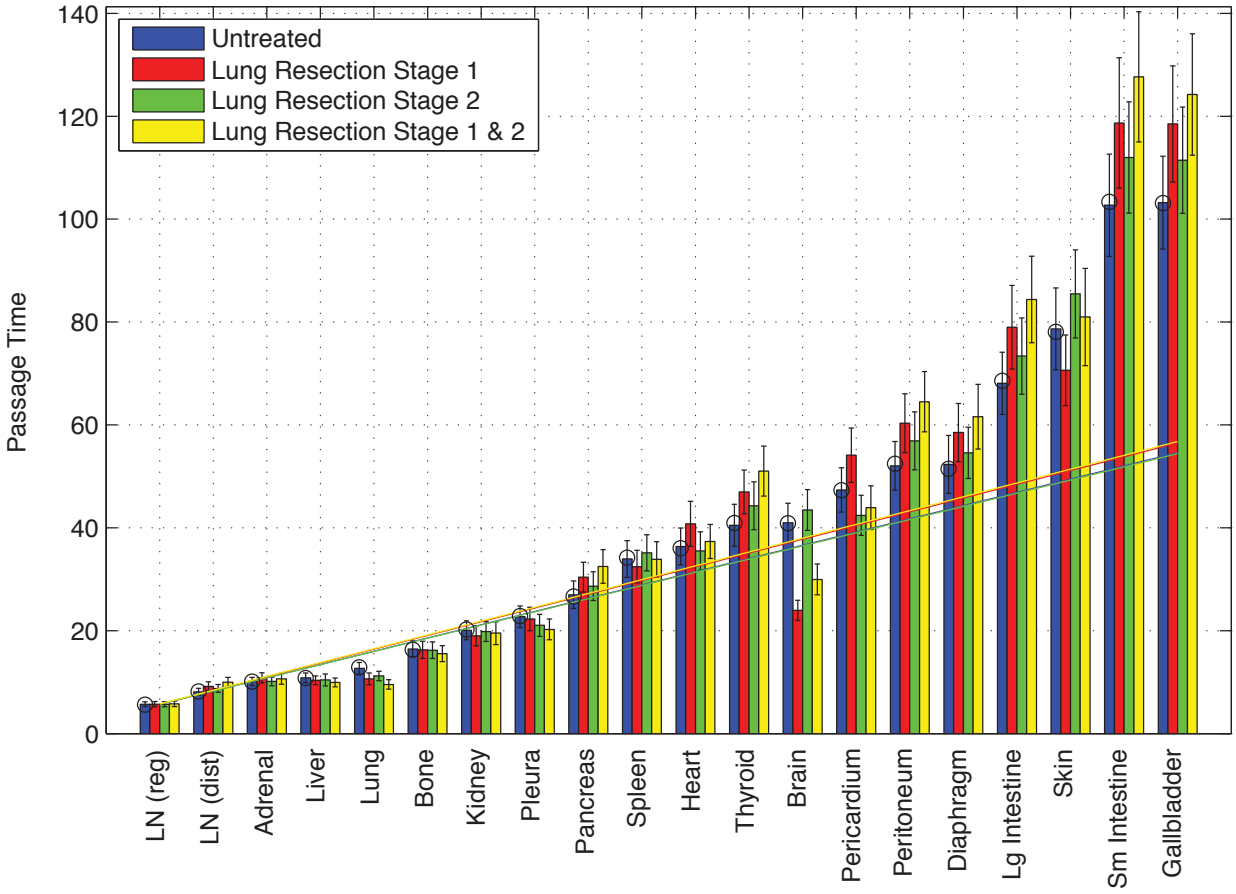


Figure 3.6: Mean first-passage times from lung to each of the metastatic sites
Blue shows the baseline (untreated population) model, red shows the baseline model with assimilated stage I resections, and green shows baseline model with assimilated stage II resections. Lines are linear curve fits to first 9 entries. Error bars show 1 SD from the mean. See text for details.

### 3.3.6 Assimilating new autopsy data of adenocarcinoma lung cancer patients undergoing complete resection

Figure 3.6 (more details are shown in Table 3.2) also shows metastatic pathways and mean first-passage times using the model with assimilated data from [73], an autopsy dataset tracking a cohort of patients with adenocarcinoma of the lung (ACL) who underwent complete lung resection. Of these, 35 survived more than 30 days after resection, 22 were classified as stage I, and 13 as stage II. We assimilated their metastatic tumor distribution from an autopsy study into our baseline (untreated population) model, recalculated the Markov transition matrix and all mean first-passage times. The results are shown in Figure 3.6 (and the middle and right columns of Table 3.2). Stage I are shown in medium dark, stage II in light gray.

Comparing the columns of Table 3.2, the main change in the spatial pathways shows up in the fifth entry down, where the Lung $\rightarrow$ Adrenal $\rightarrow$ LN (dist) pathway drops in probability on the list of the stage I treated patients but not as much as for the stage II treated patients. Lung resection seems to alter this important pathway, particularly for stage I patients, making it less likely to occur, perhaps by disruption of lymphatic connections between the primary tumor and ipsilateral adrenal gland. The overall probabilities of each of the pathways in the treated population also decrease from the untreated population.

The effect of treatment on the overall mean first-passage times is shown in Figure 3.6. The corresponding curve fit to the first 9 sites follow the same general linear trend as in the untreated population, $MFPT = a \cdot t + b$, but with $a = 2.68, b = 1.55$ (stage I, medium

dark); $a = 2.54, b = 1.91$ (stage II, light gray). The conclusions we can draw are clear: mean first-passage times increase overall with the stage I treated cohort, shown by the increase in slope over the untreated slope, but not with the stage II treated cohort. Interestingly, the MFPT back to lung in the treated cohort actually decreases with treatment. As lung is classified as a sponge in our model, this does not seem to have a negative overall effect on the general trend of increasing passage times with treatment. In contrast, the MFPT back to adrenal gland (the key spreader) with the treated cohort increases. This enhances the overall increase in MFPTs for the treated cohort. The mean first-passage times increase most in the subgroup of stage I patients, indicating that complete lung resection is more effective in this group compared with the stage II subgroup. To summarize, our model shows that lung resection for patients with ACL seems to generally increase overall MFPTs of metastases for stage I patients, and it does this by (i) altering a key pathway from lung to adrenal gland to distant lymph nodes, (ii) increasing mean first-passage times to the adrenal gland (spreader), (iii) decreasing mean first-passage times back to the lung (sponge), and (iv) reducing the overall top pathway probabilities. Lung resection seems to have very little impact on stage II patients. The failure of resection to improve metastasis-free survival in stage II patients with lung cancer may occur because the regional lymph nodes act as a sponge (Figure 3.4), potentially suppressing early metastasis when not removed. However, because the risk of local disease is high in lung cancer, surgery remains the preferred treatment in stage II disease.

## 3.4 Discussion

This chapter focused on the importance of the primary and the role it's metastatic sites played in the big picture. As a result, we looked at the top ranked two-step paths and their role in cancer being a multidirectional spreader and classified sites as spreaders and sponges to analyze how a metastasis effects disease progression. The data assimilation section shows how new data can be incorporated into the existing model without having to start at the beginning of the process and also how widely cancer can differ depending on the stage that it is in. This type of information and classification can tell a great amount about the model that is not easily seen without a bit of analysis.

Chapter 4

# Comparisons of 8 major cancer types

## 4.1 Introduction

As mentioned earlier, the entire process of creating a Markov model and analyzing it's transition probabilities and dynamics can be repeated for any primary cancer type that we have the proper data for. This section focuses on analyzing and comparing the models for 12 major cancer types (lung, breast, prostate, colorectal, pancreatic, ovarian, cervical, skin). These cancers were chosen because they are some of the most common cancers found today. We focus on the models' convergence to their steady-state, their network, multistep pathway, and reduced order diagrams, and their mean first-passage times. Included in this analysis is classification of first- and second-order sites, spreaders and sponges, and a visual 'predictability' associated with the model.

## 4.2 Network diagrams

Just as with primary lung cancer, each metastatic model is built from the dataset listed in [19] by iterating upon an initial guess matrix that is constructed from the overall cancer distribution and it's respective distribution. An ensemble of matrices is calculated for each and used for comparison for the rest of this chapter. The first, and most obvious thing to look at is the structure of each network, which is shown in Figure 4.1. From a global standpoint, these look practically identical.

Closer inspection and an analysis of the strengths of the connections (not shown in figures) shows more subtle differences. The skin network is composed of the most nodes (30) while the prostate network has the fewest (21). Despite this, with roughly half the number of connections, the prostate network's strongest connection is more than twice that of the strongest connection in skin, and 5.77 times stronger than the average weighting of it's network (compared to 3.93 for skin). The ovarian network boasts the largest maximum weight connection along with the largest maximum-average ratio right at 9.05 The other networks, excluding skin and prostate, range from 26-28 nodes and 676-784 connections between them. This is a perfect indication of how similar cancer appears from a global standpoint, but a more detailed look can reveal how different each can be.

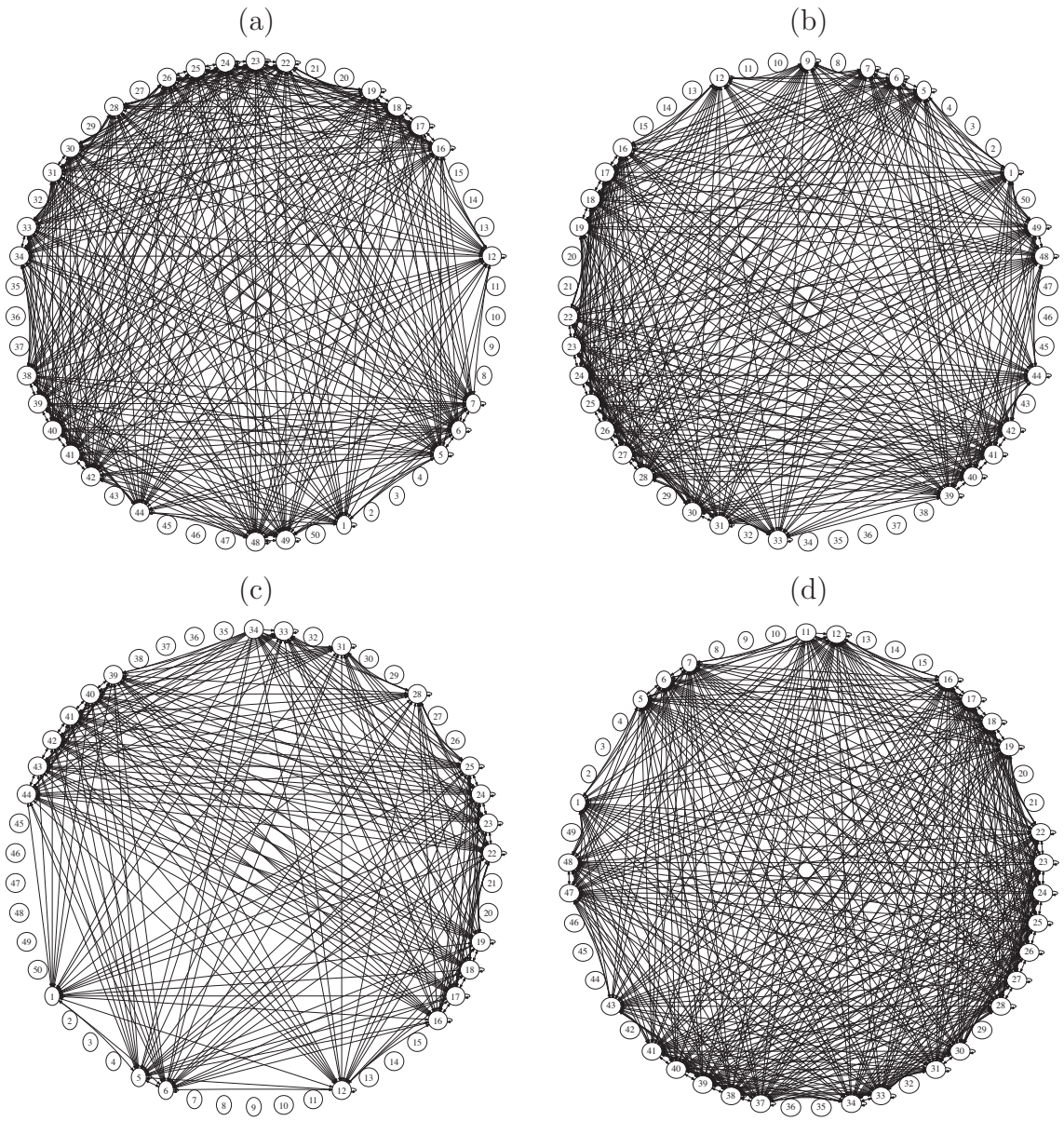


Figure 4.1: Converged cancer networks of 8 cancer types

Converged cancer networks shown as circular, bi-directional, weighted graphs. Arrow heads placed on the end or ends of the edges denote the direction of the connections. Edge weightings are not shown. Primary cancer placed on top. a. Primary lung cancer. b. Primary breast cancer. c. Primary prostate cancer. d. Primary colorectal cancer. (*Continued on the following page.*)

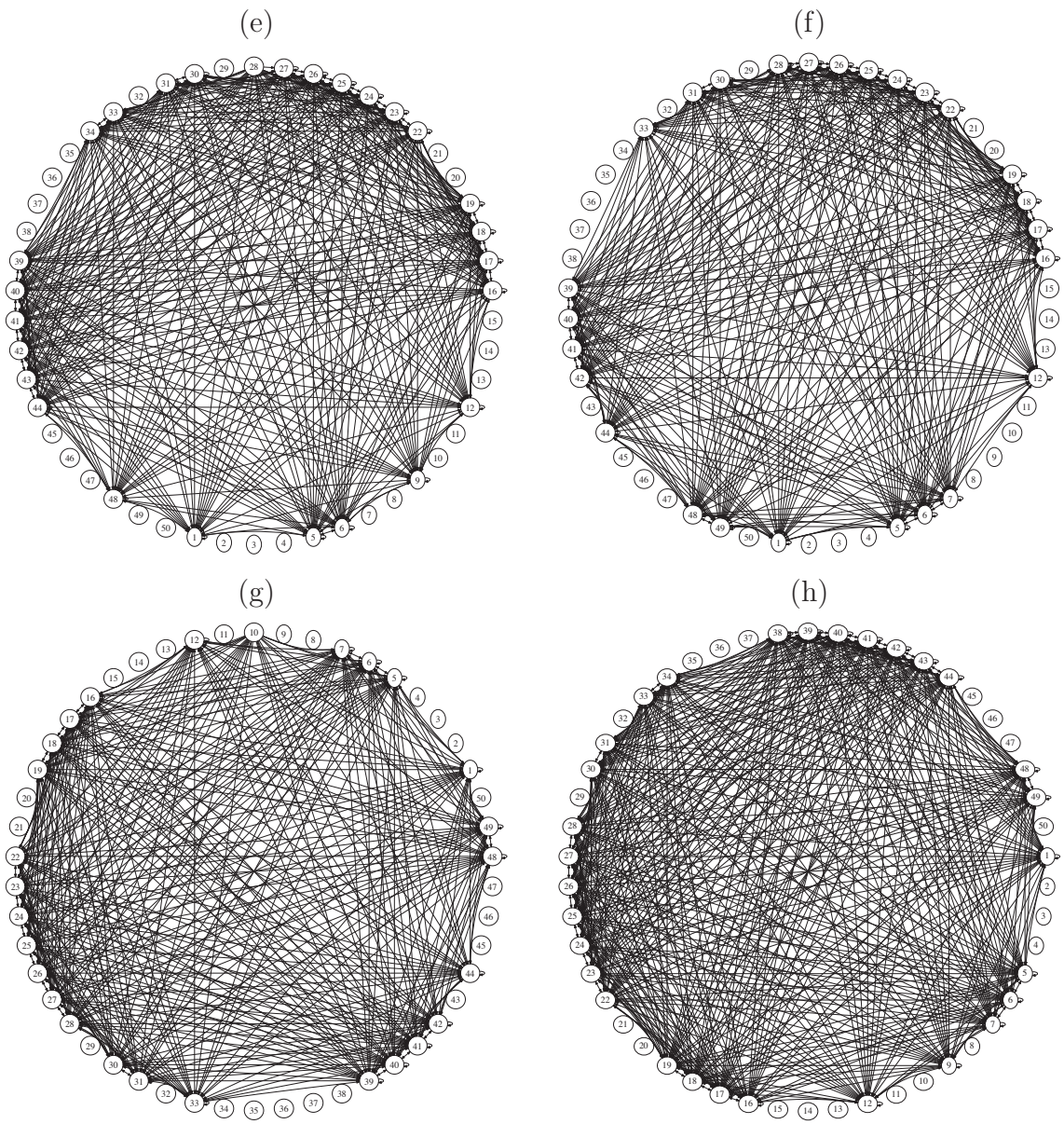


Figure 4.1 (*continued.*)

Converged cancer networks shown as circular, bi-directional, weighted graphs. Arrow heads placed on the end or ends of the edges denote the direction of the connections. Edge weightings are not shown. Primary cancer placed on top. e. Primary pancreatic cancer. f. Primary ovarian cancer. g. Primary cervical cancer. h. Primary skin cancer.

## 4.3 Convergence to steady-state

As with every model that is created, verification that it is performing properly is essential.

The most logical way to do this with these metastatic cancer models is to ensure that each

progresses to it's steady-state. Figure 4.2 shows the dynamical progression of each initial-state vectors plotted on a semi-log plot as the Euclidian norm between the current state vector and the steady-state ($\|\vec{v}_k - \vec{v}_\infty\|$). A line of best fit indicates exponential progression in the form of

$$\|\vec{v}_k - \vec{v}_\infty\| \sim \alpha \exp(-\beta k) \tag{4.1}$$

As shown in the figure, the $\beta$ values are roughly similar ranging from $-3.5$ to $-5.4$. In fact, three pairs, colorectal and ovaries, breast and prostate, and lung and skin, have nearly identical decay rates differing by at most 0.05. This would indicate that, from a global view, each cancer model behaves more or less the same in terms of overall disease progression. The same dynamical progression is shown again in Figure 4.3 on a linear plot. This more clearly shows the exponential decay, but more importantly, it shows the convergence of each model to each steady-state in approximately 2 steps. This indicates that in each model, it is indeed the first two steps in the progression that are the most relevant.

## 4.4 Multistep Pathway diagrams

With the multistep pathway diagrams, shown in Figure 4.4, its much easier to see differences between the networks. The ovarian and skin networks are the only two models that do not contain second-order nodes. This indicates that every connection emanating from the primary is equally large or greater than if it went through an intermediate site. On the other

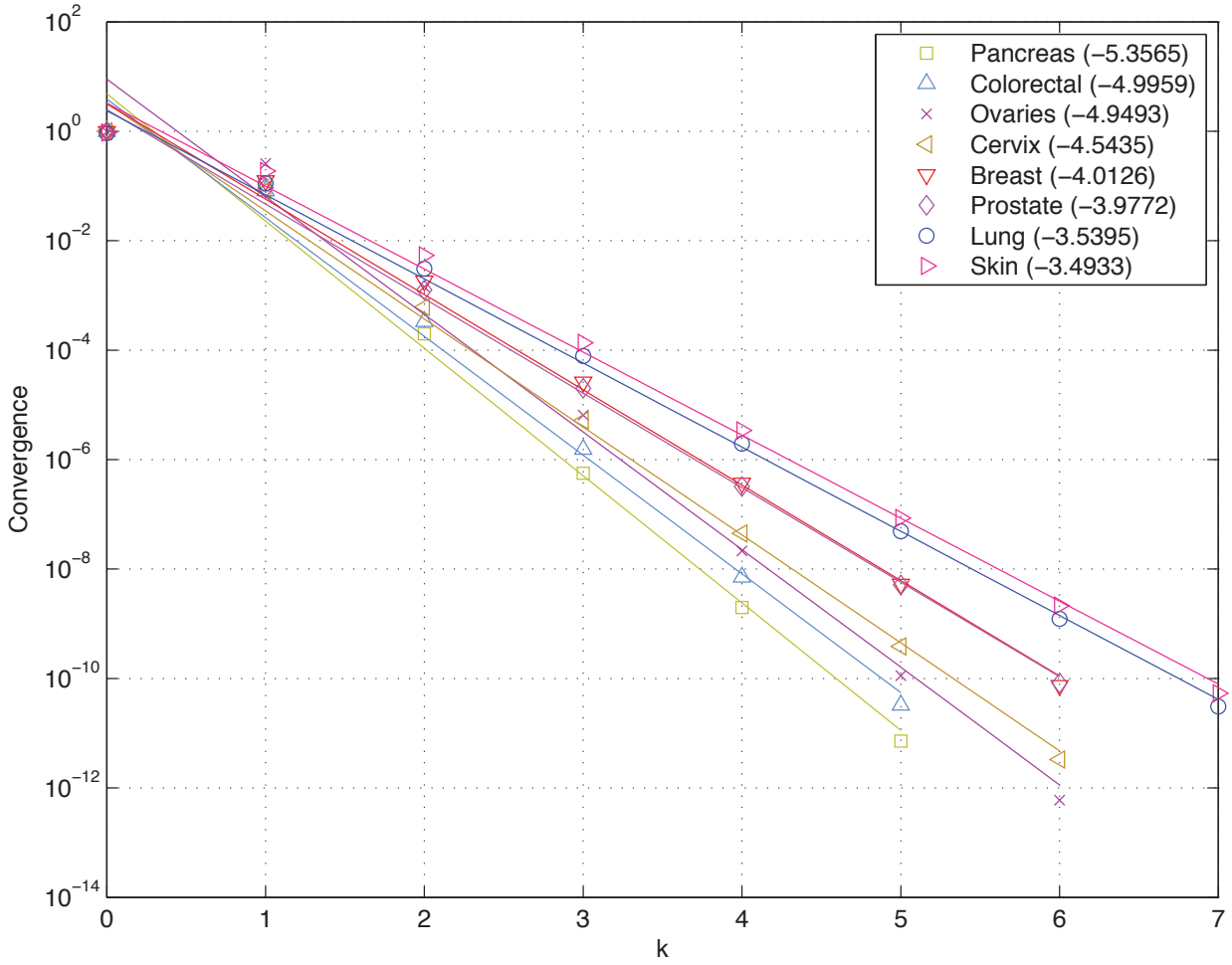


Figure 4.2: Dynamical progression of $\vec{v}_0$ of 8 cancer types (semi-log)
Dynamical progression of initial-state vectors plotted on a semi-log plot. Values plotted are Euclidian norm ($\|\vec{v}_k - \vec{v}_\infty\|$) between state vector and steady-state. Line of best fit follows exponential decay in the form $\sim \alpha \exp(-\beta k)$. $\beta$ values for each primary listed in legend.

hand, the models for prostate, colorectal, pancreatic and cervical cancers do not illustrate primary reseeding nor primary self-seeding. Evidence of this is found in the original dataset in the lack of metastases formed at other locations in the primary organ.

In most of the networks, the regional and distal lymph nodes have some of the strongest connections to the primary. The exceptions to this are the skin and ovarian network. Even

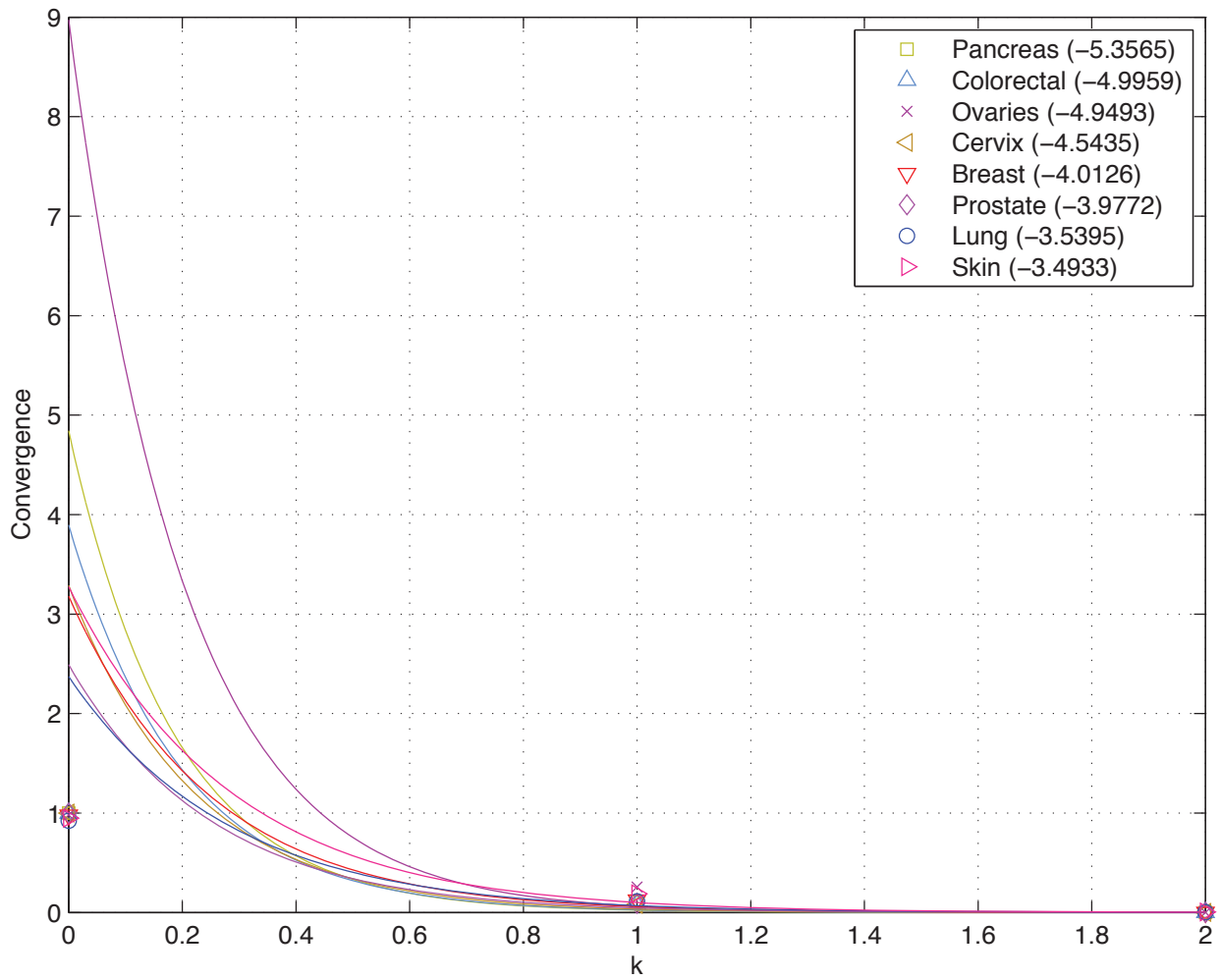


Figure 4.3: Dynamical progression of $\vec{v}_0$ of 8 cancer types (linear)
Dynamical progression of initial-state vectors plotted on a linear plot. Values plotted are Euclidian norm ($\|\vec{v}_k - \vec{v}_\infty\|$) between state vector and steady-state. Line of best fit follows exponential decay in the form $\sim \alpha \exp(-\beta k)$. $\beta$ values for each primary listed in legend. Convergence is nearly complete after 2 steps.

though the distal lymph nodes forms the second strongest connection with skin, the regional lymph nodes falls down to position 12. Another important observation is the location of skin's self-seeding loop at position 3. This is unusually high as compared to the other

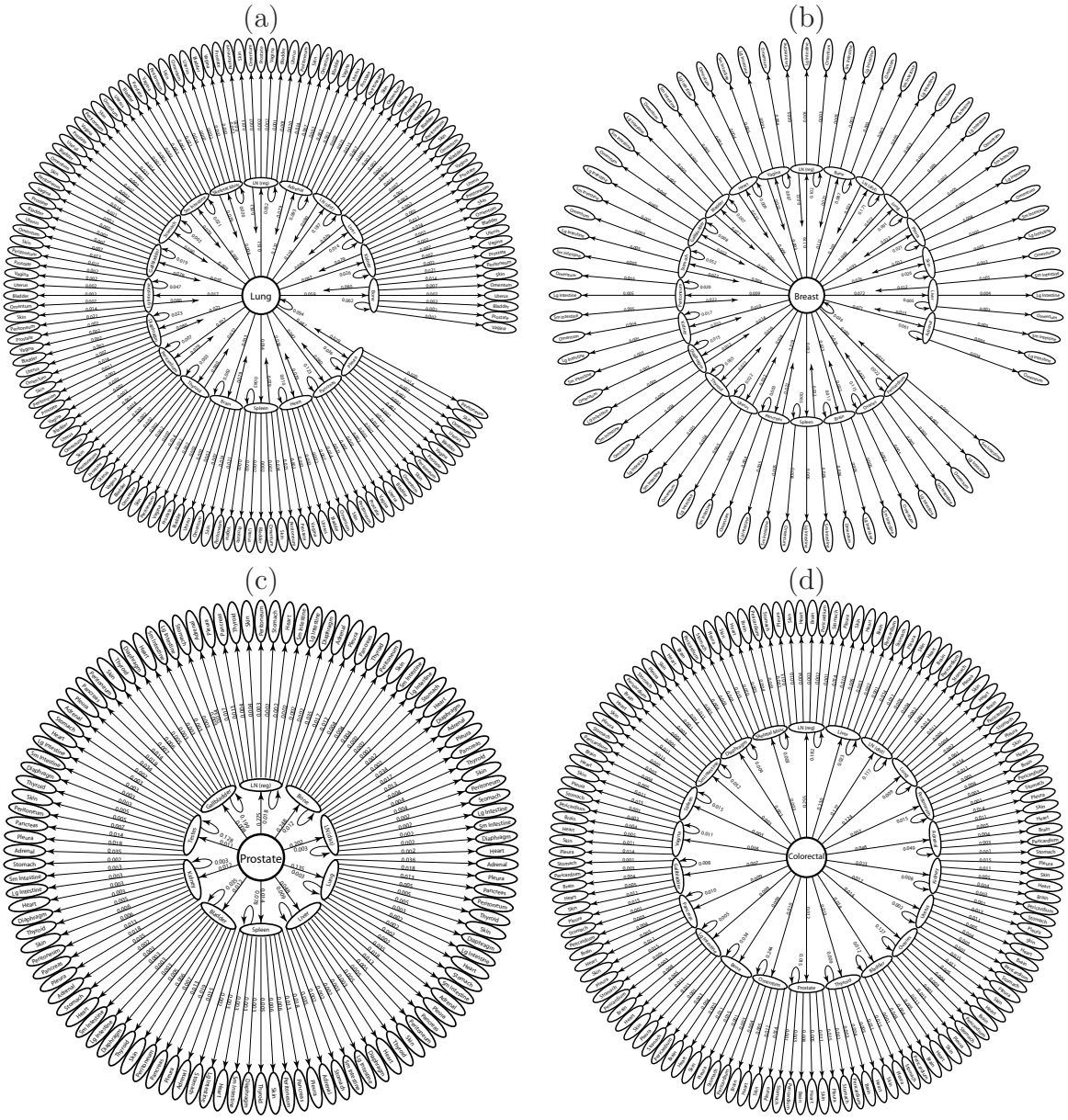


Figure 4.4: Multistep pathway diagrams of 8 cancer types

Ensemble averaged one-step pathway diagram. Primary tumor is at the center, next ring out are the first-order sites showing their direct connection from the lung, with transition probabilities getting weaker in clockwise direction. Next ring out are the second-order sites and their connections from the first-order sites. a. Primary lung cancer. b. Primary breast cancer. c. Primary prostate cancer. d. Primary colorectal cancer. (*Continued on the following page.*)

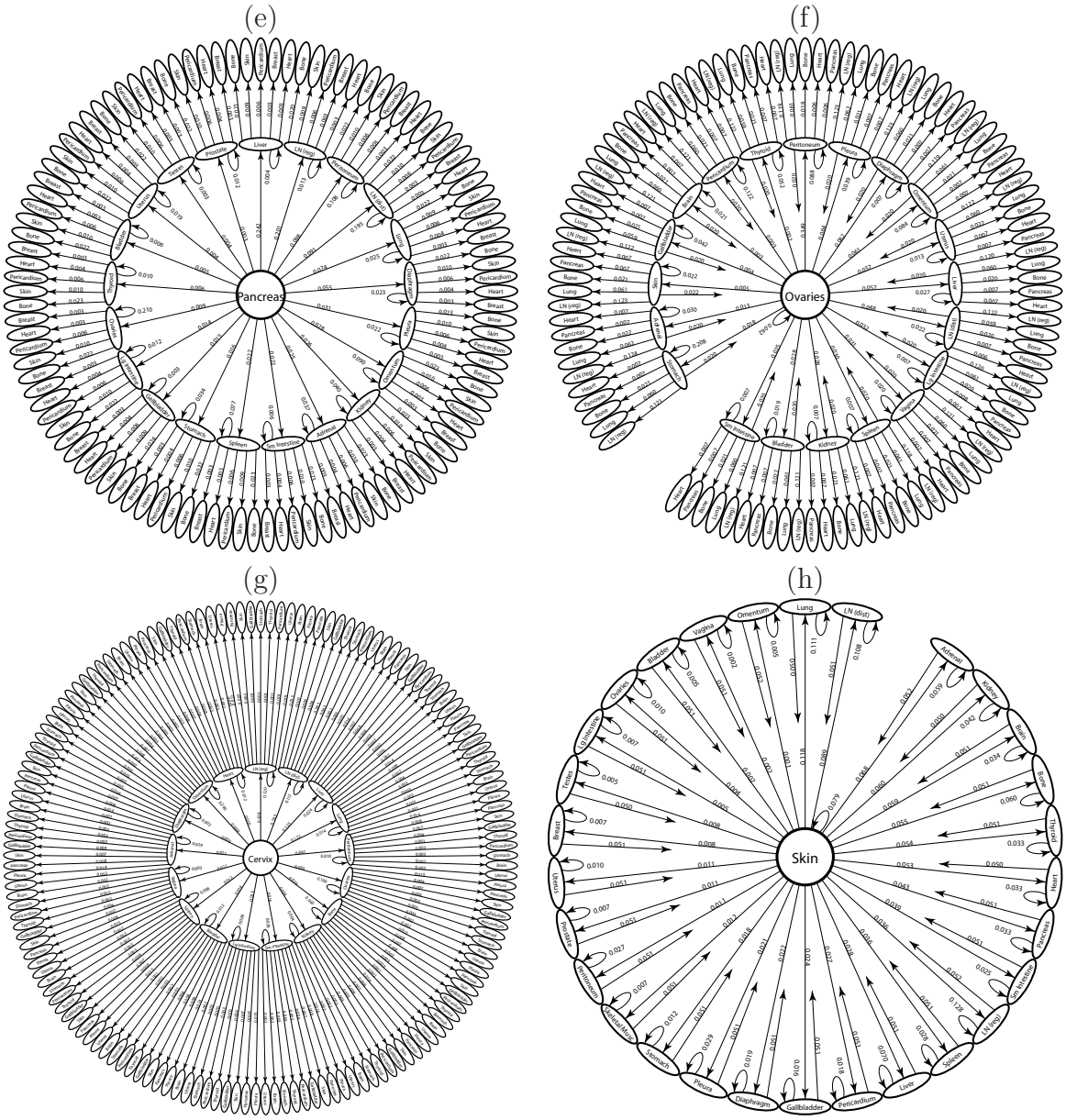


Figure 4.4 (*continued.*)

Ensemble averaged one-step pathway diagram. Primary tumor is at the center, next ring out are the first-order sites showing their direct connection from the lung, with transition probabilities getting weaker in clockwise direction. Next ring out are the second-order sites and their connections from the first-order sites. e. Primary pancreatic cancer. f. Primary ovarian cancer. g. Primary cervical cancer. h. Primary skin cancer.

models that fall at 7, 8, and 9. This may indicate that skin more easily and readily spreads throughout it's primary organ unlike other cancers.

## 4.5 Reduced diagrams

The natural progression of the model analysis leads to the creation of the reduced pathway diagrams and the classification of spreaders and sponges. Figure 4.5 shows the reduced diagrams for the 8 cancer models with their highlighted spreaders and sponges. Once again, these diagrams are constructed from the top 30 two-step pathways emanating from the primary cancer. These values are then used to calculate the probability in and out of each node to classify the spreaders and sponges. The percentage located below the primary indicates the total percentage that the 30 paths correspond to out of all the two-step paths. This percentage will be addressed later in the section.

Looking at the reduced diagrams, the most obvious difference between the models is the number of nodes involved. The lung, breast and pancreas models only use 8 of their nodes, while prostate and colorectal use 7, ovarian and cervical use 9, and skin uses 10. Regardless of the number of nodes, the most 'active' node is the first one located at 12:00. Seeing as this is contains the strongest connection to the primary, this makes sense in terms of it having a good amount of two-step pathways in the top 30. The next most obvious thing is the multidirectional pathways associated with each model. While all of them show metastasis reseeding (yellow loops attached to metastatic nodes) in their top 30 paths, only lung,

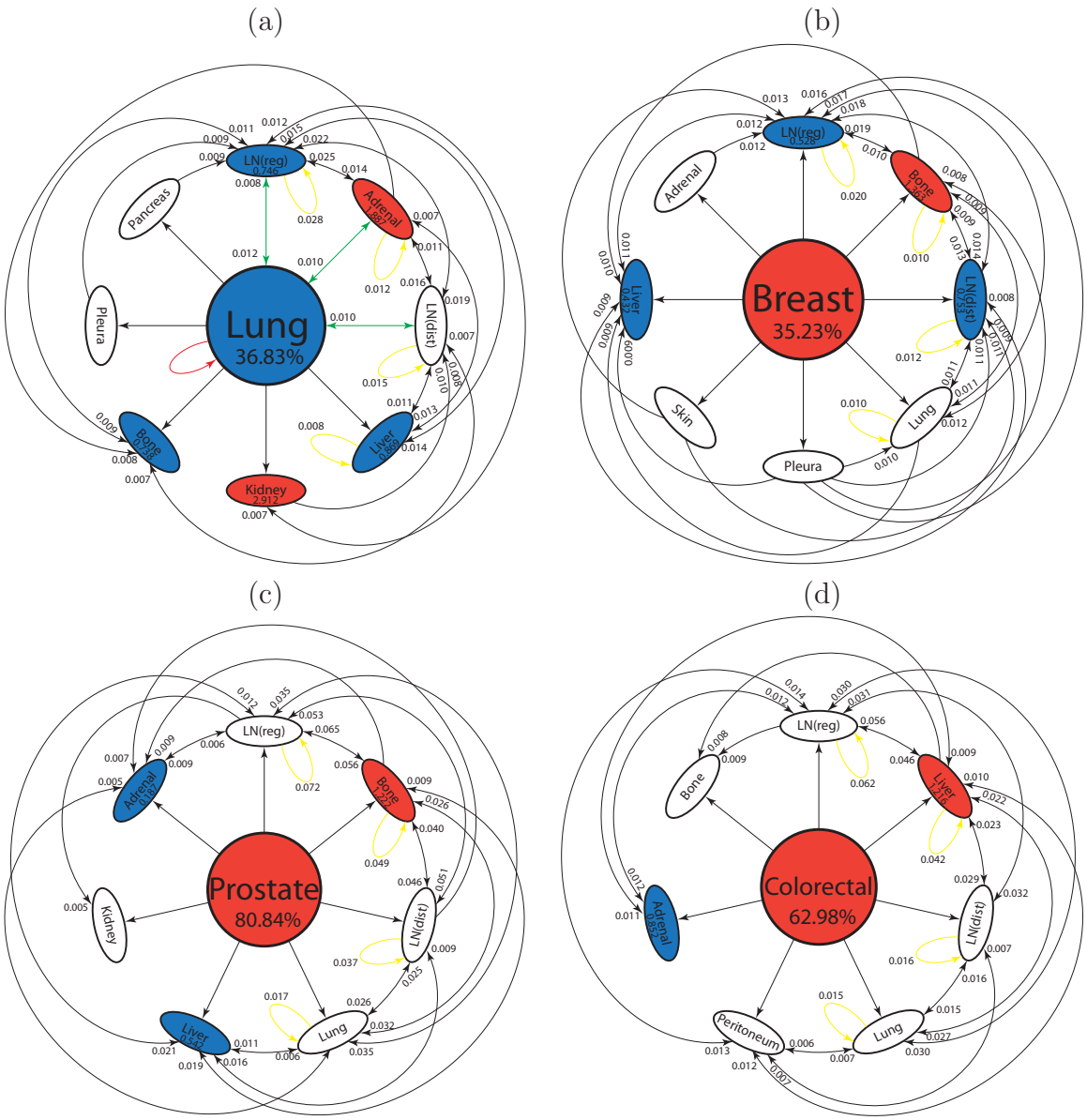


Figure 4.5: Reduced pathway diagrams of 8 cancer types showing top 30 paths
Top 30 two-step pathways emanating from primary tumors (total pathway probability listed in center node), obtained by multiplying the edges of the one-step edges comprising each two-step path. Edges without numbers are one-step paths. All other numbered edges mark the second edge in a two-step path, with numbers indicating the two-step probabilities. a. Primary lung cancer. b. Primary breast cancer. c. Primary prostate cancer. d. Primary colorectal cancer. (*Continued on the following page.*)

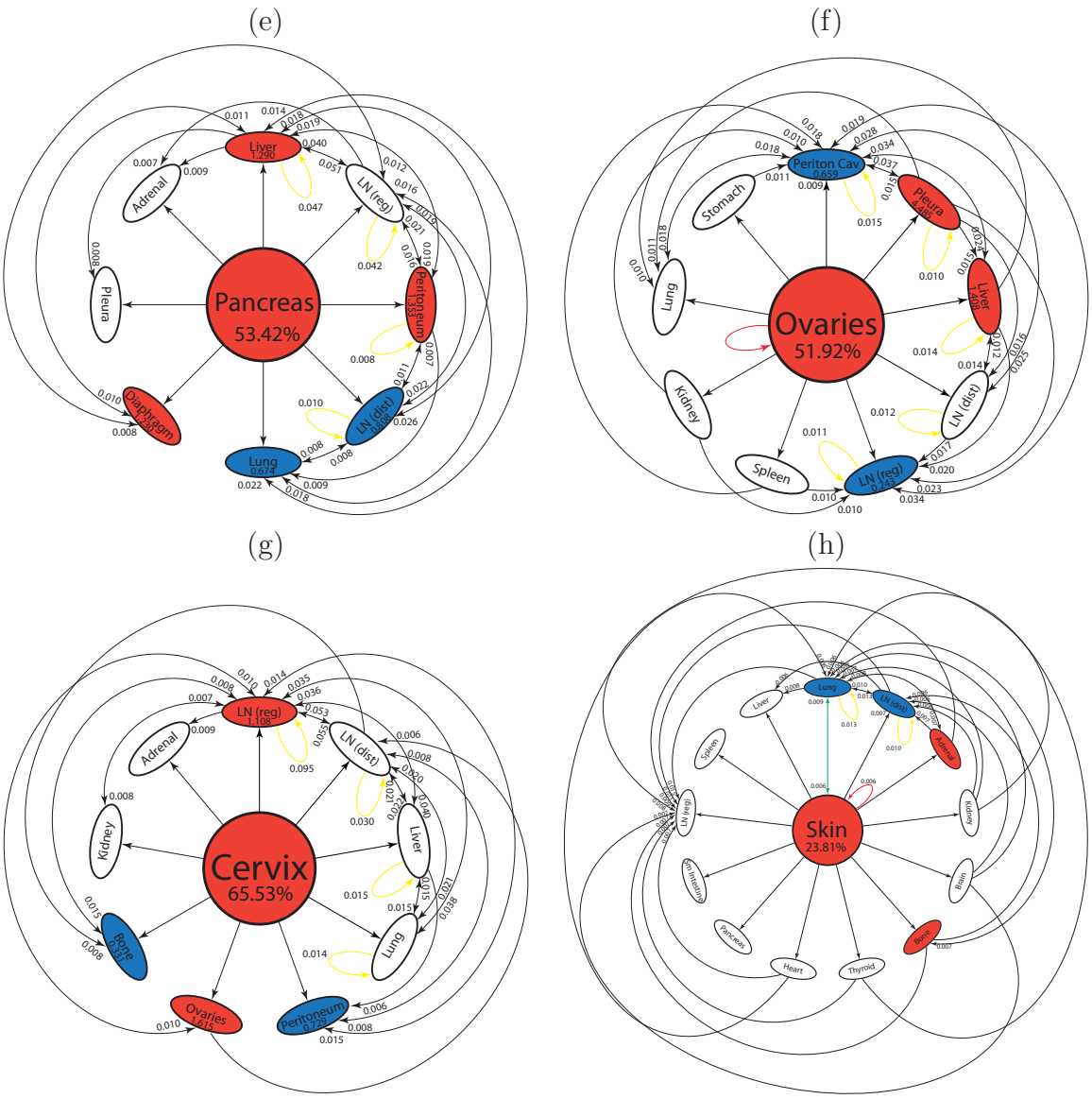


Figure 4.5 (*continued.*)

Top 30 two-step pathways emanating from primary tumors (total pathway probability listed in center node), obtained by multiplying the edges of the one-step edges comprising each two-step path. Edges without numbers are one-step paths. All other numbered edges mark the second edge in a two-step path, with numbers indicating the two-step probabilities. e. Primary pancreatic cancer. f. Primary ovarian cancer. g. Primary cervical cancer. h. Primary skin cancer.

ovarian, and skin models show primary self-seeding (red loops attached to the primaries), and only lung and skin show primary reseeding (green arrows pointing back to the primary).

### 4.5.1 Spreaders and sponges

One of the most important pieces of information to gather from these reduced diagrams is the classification of the spreaders and sponges. Since these tell us what spreads and what absorbs the CTCs, they are very important to analysis of the models. The first key observation is that the number of each differs from primary to primary. While lung and pancreas have five stand-out sights each, the colorectal model has only two. Most of the models have a fairly equal number of spreaders and sponges (differing at $\pm 1$), yet breast has three sponges to it's singular bone spreader.

Another noteworthy difference between the models is how a node acts like a sponge in one model and a spreader in another. A perfect example of this is bone in breast, prostate, and skin cancers, which acts like a spreader, yet acts like a sponge in lung and cervical cancers. Another major example is the liver acting like a sponge in lung, breast, prostate, and ovarian cancer, and a spreader in colorectal and pancreatic. This shows how different cancer can act on a smaller scale. The same sites are major players in the disease progression, but for different reasons depending on which primary the metastasis came from.

### 4.5.2 Two-step pathway percentages

Looking back at the percentage of two-step paths that these 30 paths take up, we begin to see a different picture. These values range from 23.81% to 80.84% and give an indication of how complex and unpredictable each cancer really is. On one end, prostate's 30 paths representing 80% of all it's paths, says that this diagram is a good representation of how prostate cancer metastasizes throughout the body and that this process is fairly predictable. On the other end of the spectrum, skin's 30 paths representing only 24%, indicates that the exact opposite, that this is not a good representation of skin cancer metastasizing and that the process can be very unpredictable.

Another way of looking at these diagrams is to normalize them so all of them are showing approximately the same percentage instead of the same number of paths. Figure 4.6 shows this new approach and how easy it is to see the predictability in the models. The percentage for the figures is fixed at $\sim 35\%$, and the appropriate connections for each is adjusted. While breast remains the same and lung only loses 2 of it's connections, the other models change much more drastically. Ovarian falls to 18 paths, pancreatic to 12, colorectal to 9, cervical to 7, and prostate to 6. The biggest change, is that skin has to nearly double it's two-step pathways (54) in order to reach the 35% value.

This shows perfectly how predictable each cancer model is. The prostate, colorectal, and cervical models contain most of their metastatic progression information in only a handful of connections between a few sites. conversely, the skin model is very unpredictable in the sense that the same amount of information is spread across many more pathways. While

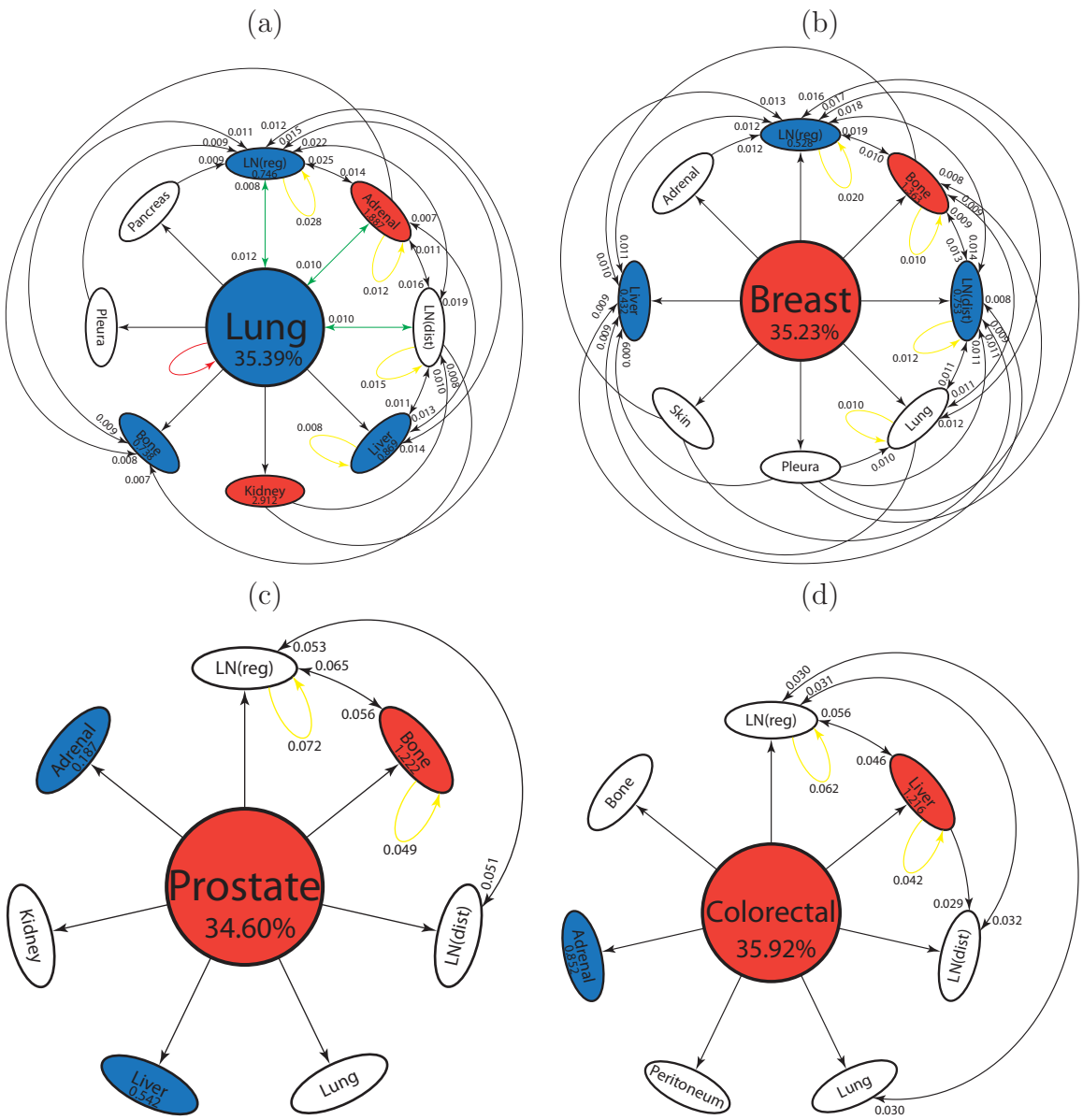


Figure 4.6: Reduced pathway diagrams of 8 cancer types showing top 35%
Top two-step pathways totaling $\sim$ 35% emanating from primary tumors (actual pathway probability listed in center node), obtained by multiplying the edges of the one-step edges comprising each two-step path. Edges without numbers are one-step paths. All other numbered edges mark the second edge in a two-step path, with numbers indicating the two-step probabilities. a. Primary lung cancer. b. Primary breast cancer. c. Primary prostate cancer. d. Primary colorectal cancer. (*Continued on the following page.*)

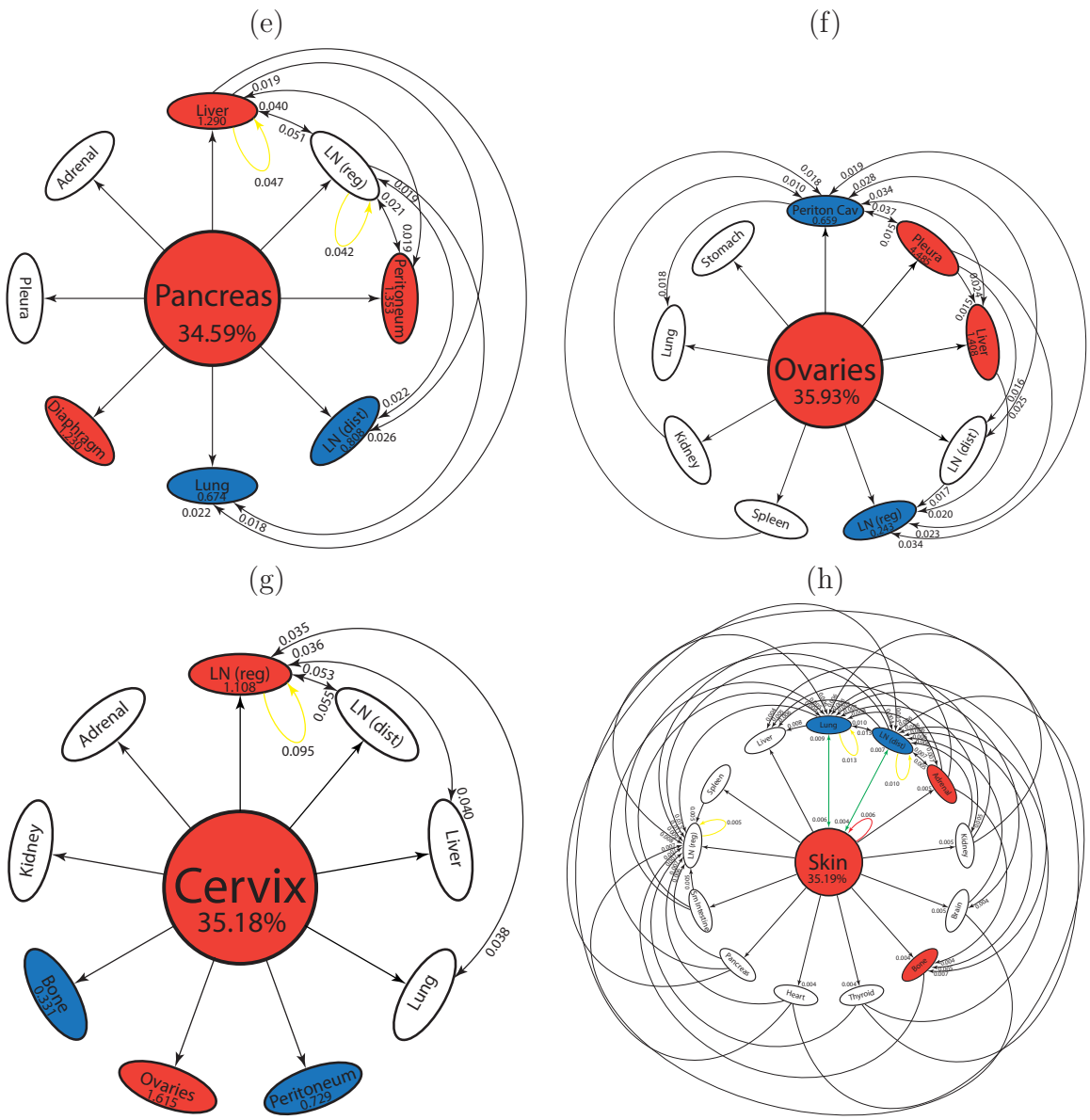


Figure 4.6 (*continued.*)

Top two-step pathways totaling $\sim$ 35% emanating from primary tumors (actual pathway probability listed in center node), obtained by multiplying the edges of the one-step edges comprising each two-step path. Edges without numbers are one-step paths. All other numbered edges mark the second edge in a two-step path, with numbers indicating the two-step probabilities. e. Primary pancreatic cancer. f. Primary ovarian cancer. g. Primary cervical cancer. h. Primary skin cancer.

the other models fall somewhere in between these two extremes, skin and prostate serve as benchmarks in terms of how predictable each model is as compared to one another.

## 4.6 Mean first-passage times

Upon analyzing the mean first-passage times of each model, more similarities and differences begin to emerge. Figure 4.7 shows the graphical representation of this along with standard deviation bars and markers (listed as 'o's) for the analytical values. By looking at the shape of the graphs, we can group them into two different categories: i.) a gradual progression of passage times and ii.) distinct groups of progression times. The prostate, pancreatic, and ovarian models fall in the second category while the rest follow the gradual progression. This is indicating that prostate, pancreatic, and ovarian cancers may metastasize to some sites, lay dormant for a while, and then continue metastasize to additional sites. The other models show a more consistent metastatic pattern that stays active throughout the disease.

The next easily identifiable difference is the range of progression times. Since the first site is metastasized to a value of 1, it is easy to compare the difference between the first and last sites' metastasis times. While the ovarian network can fully metastasize in as little as 40 time steps, it takes the cervical model nearly 7 times the amount of time steps. This says that all of the sites in the ovarian model metastasize at a more uniform rate as compared to the cervical model. In order to achieve this same uniformity, the cervical model would have to be truncated after the skin node (the $20^{th}$ largest passage time). Although this is no indication of how long the cancer will actually take to metastasize, because these are

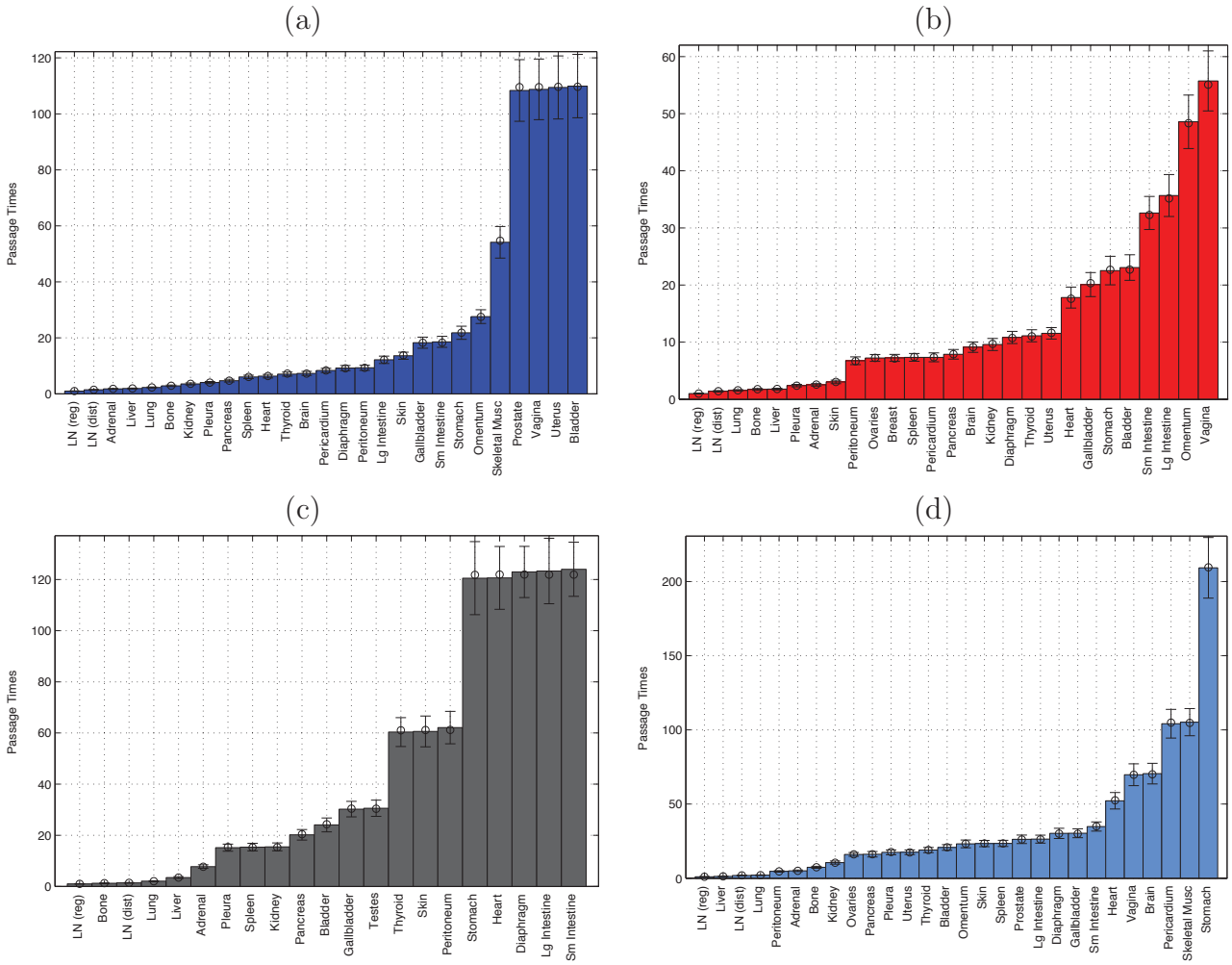


Figure 4.7: Mean first-passage time histograms of 8 cancer types showing top 35% Mean first-passage time histogram for Monte Carlo computed random walks all starting from primary tumor. Error bars show one standard deviation. Values are normalized so that first passage time has value 1, and all others are in these relative time units. a. Primary lung cancer. b. Primary breast cancer. c. Primary prostate cancer. d. Primary colorectal cancer. (*Continued on the following page.*)

arbitrary, model based times, it is good to know how spread out the metastatic cascade can occur on.

A more detailed inspection of these graphs will reveal a different picture. While all of the models indicate that the regional and distal lymph nodes, liver and lung are metastasized

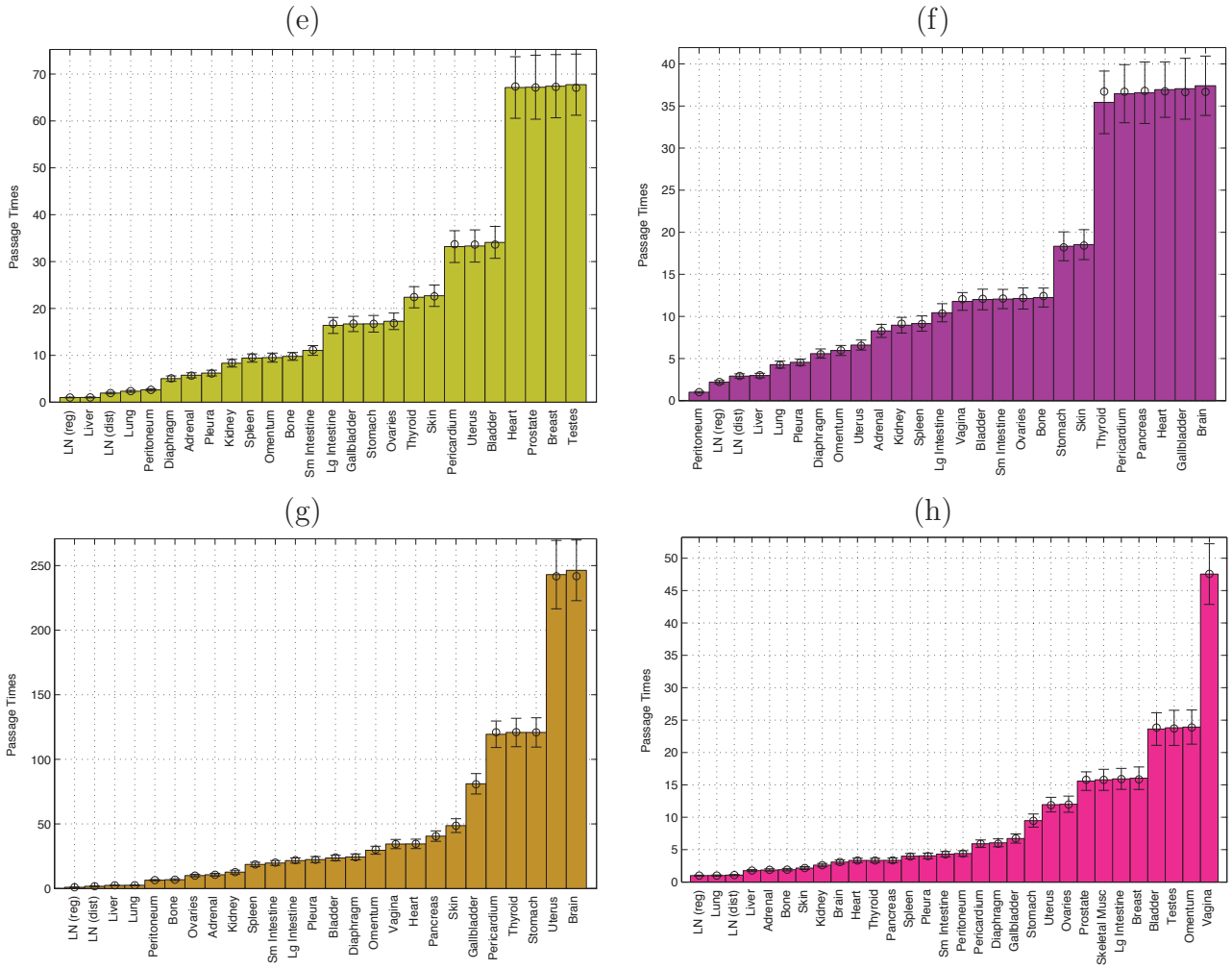


Figure 4.7 (*continued.*)

Mean first-passage time histogram for Monte Carlo computed random walks all starting from primary tumor. Error bars show one standard deviation. Values are normalized so that first passage time has value 1, and all others are in these relative time units. e. Primary pancreatic cancer. f. Primary ovarian cancer. g. Primary cervical cancer. h. Primary skin cancer.

to very soon in the disease progression, there are some nodes that are unique to each cancer that are equally as fast. The lung model boasts a very important adrenal gland located at position #3. This site is uniquely important to lung cancer, but not a major player in the other models. In fact, the highest passage time location that is reached by the adrenal gland

outside of the lung model is #5 in skin cancer. Likewise for ovarian cancer with peritoneum and prostate cancer with bone. These important sites indicate, once again, that even though each cancer is globally the same, there lies subtle differences on a smaller, more personal scale.

## 4.7 Discussion

This chapter shows how easy it is to analyze and compare different cancer models in a very robust and sensible way. Not only do we analyze the networks and their properties, but also the analytical progression of the disease and the classification of the sites. This shows us how predictable some cancers are and also how uniform the metastatic progression. By performing such analyses on mathematical models, it is much easier to compare similarities and differences between them than in living people. In the long run, and with patient specific models, these models have the potential to save money, time, and lives.

Chapter 5

# The entropy of metastatic cancer

## 5.1 Introduction

This chapter focuses on the entropy associated with the Markov models. The entropy is based on the distributions gathered from the dataset that the models are built from, which is also the steady-state of the model itself. After analyzing these values individually, we calculate the relative entropy between the distributions to see how they compare to general cancer. This will allow for further comparisons of cancers and also more predictions of how they will behave throughout the lifetime of the disease.

## 5.2 Methods

### 5.2.1 Brief summary of autopsy dataset used

We re-analyze the DiSibio and French [19] dataset of metastatic tumor distributions based on autopsy studies collected for 3827 untreated cadavers from 5 different cancer facilities in New

England between 1914-1943. The data reflect 9484 distinct metastatic tumors distributed over 30 anatomical sites for all of the major tissue cancers. The data represent 'natural' disease progression, which is useful, but we caution that brain metastases are under-represented in the data since examination of the intracranial contents at that time was not routinely performed. The data has been used in [54] to develop a Markov chain model for lung cancer progression, where the autopsy data is used as the Markov chain steady-state, from which transition probabilities are calculated. In this paper, we directly characterize the data, shown in Figure 5.1 (all cancers) and Figure 5.2 (12 different primary cancers) in terms of their empirical distributions, which predominantly follow power-law form [49]. Other related work focusing on the development of dynamical models based on metastatic progression patterns includes references [10, 32, 36, 55]. While notions of entropy have been used previously in the context of gene expression profiles and epidemiology [47, 66, 69, 75], we know of no previous work that uses these notions to characterize the complexity of large-scale progression patterns.

### 5.2.2 Definition of entropy

The notion of entropy we use is borrowed from the field of information theory and statistical mechanics [13, 37, 40, 71]. Given a probabilistic distribution of states $\vec{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ spread over N sites, the entropy associated with the distribution is given by the quantity $H_N = -\sum_{i=1}^{N} \sigma_i \ln(\sigma_i)$ where $0 \leq H_N \leq \ln N$. There are two factors that lead to increased entropy: (i) the larger the number N of sites over which the disease is distributed, the larger

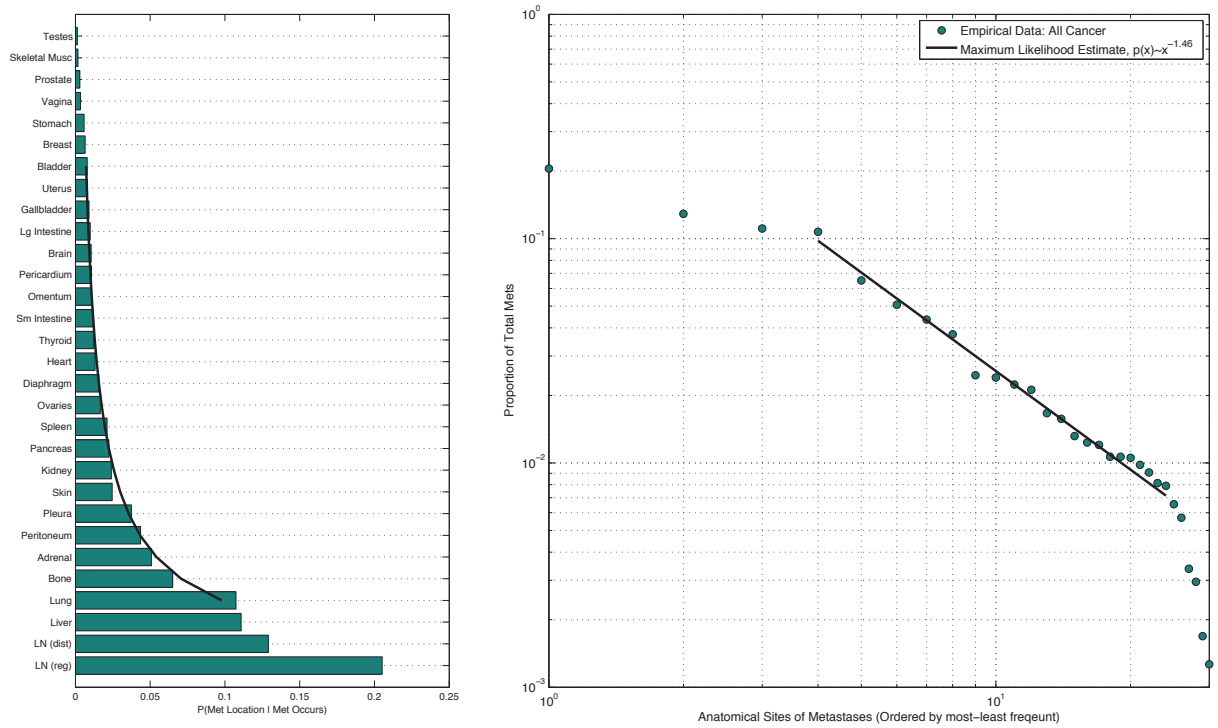


Figure 5.1: Distribution histogram and log-log plot of all cancer distribution
Histogram (left) of distribution of metastatic tumors over all cancer types from 3827 patients, 9484 metastatic tumors distributed over 30 anatomical sites. Data is plotted on log-log plot (right) showing power law form $p(x) \sim x^{-1.46}$ obtained using a maximum likelihood estimator and goodness-of-fit criteria to obtain the best range of the power law distribution [49].

the entropy; (ii) the more even the probabilities are distributed among those sites, the larger the entropy. Thus, the lowest entropy state, given by $H_N = 0$, corresponds to the distribution $\sigma_k = 1, \sigma_i = 0 (i \neq k)$. Since the probability of site 'k' being occupied is 1 and the probability of sites $i \neq k$ being occupied is 0, this state is associated with predictive certainty. In the language of statistical thermodynamics [40], this would be called a completely 'ordered' state. By contrast, the highest entropy state corresponds to the uniform distribution in which each site is equally probable, hence $\sigma_i = \frac{1}{N}, i = 1, \ldots, N$. This uniform distribution gives rise to

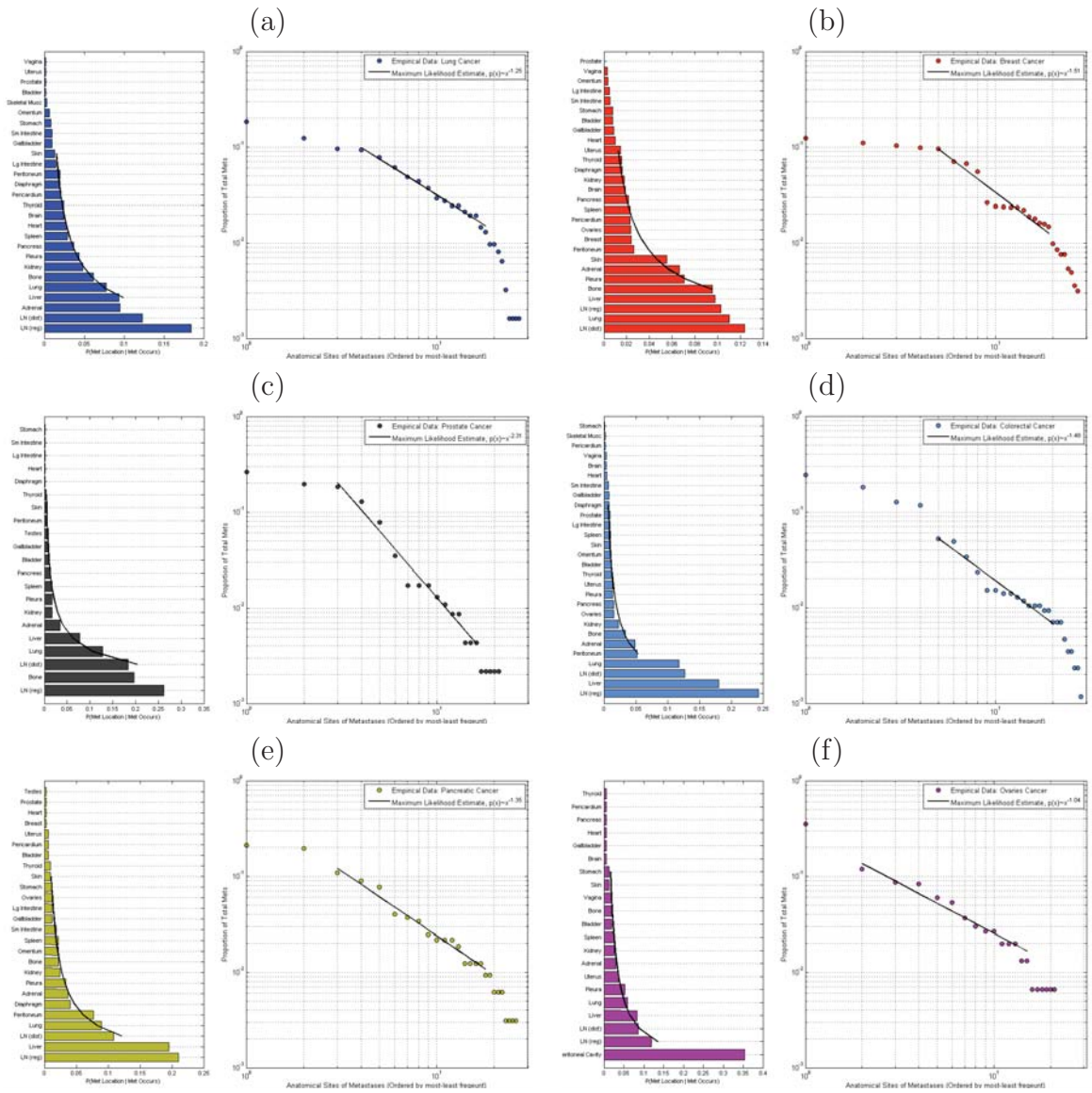


Figure 5.2: Distribution histograms and log-log plots of 12 cancer types
Histograms of distribution of metastatic tumors for primary cancers. Data is plotted on a log-log plot for each, showing power law form. (a) Lung cancer: 163 patients, 619 metastases, 27 anatomical sites, $p(x) \sim x^{-1.25}$; (b) Breast: 432 patients, 2235 metastases, 28 anatomical sites, $p(x) \sim x^{-1.51}$; (c) Prostate: 193 patients, 462 metastases, 21 anatomical sites, $p(x) \sim x^{-2.31}$; (d) Colorectal: 560 patients, 859 metastases, 28 anatomical sites, $p(x) \sim x^{-1.46}$; (e) Pancreatic: 109 patients, 323 metastases, 26 anatomical sites, $p(x) \sim x^{-1.35}$; and (f) Ovarian cancers: 86 patients, 302 metastases, 26 anatomical sites, $p(x) \sim x^{-1.05}$. (*Continued on the following page.*)

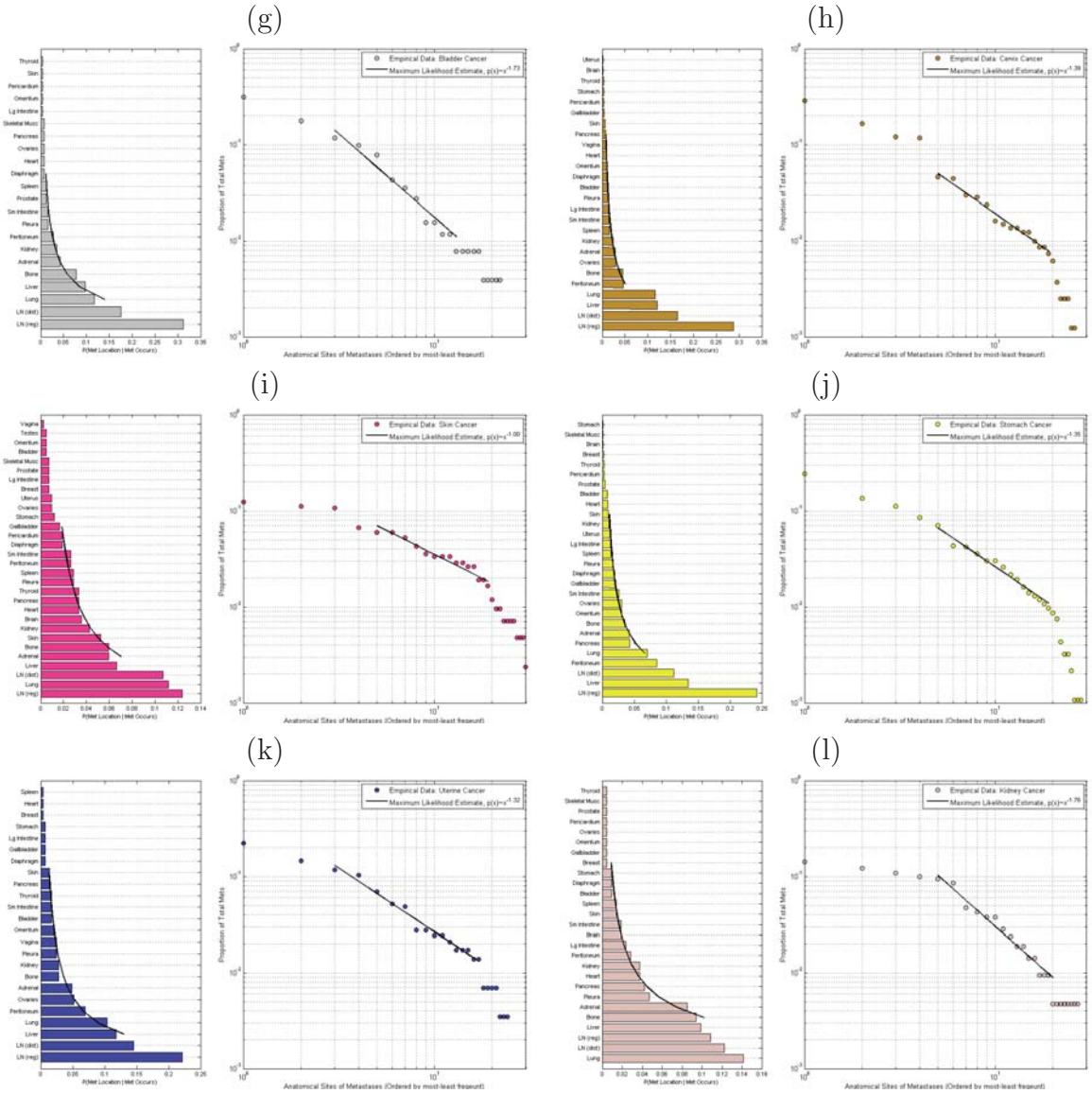


Figure 5.2 (*continued.*)

Histograms of distribution of metastatic tumors for primary cancers. Data is plotted on a log-log plot for each, showing power law form. (g) Bladder cancer: 183 patients, 256 metastases, 22 anatomical sites, $p(x) \sim x^{-1.73}$; (h) Cervical: 418 patients, 806 metastases, 26 anatomical sites, $p(x) \sim x^{-1.39}$; (i) Skin: 161 patients, 420 metastases, 30 anatomical sites, $p(x) \sim x^{-1.00}$; (j) Stomach: 348 patients, 928 metastases, 28 anatomical sites, $p(x) \sim x^{-1.35}$; (k) Uterine: 120 patients, 289 metastases, 24 anatomical sites, $p(x) \sim x^{-1.32}$; and (l) Kidney cancers: 62 patients, 212 metastases, 26 anatomical sites, $p(x) \sim x^{-1.76}$.

a maximal entropy value of $H = \ln N$. For this distribution, since each site is equally likely to occur with probability $1/N$, the predictive certainty associated with this distribution is minimal, yielding the highest possible entropy value. We note that the entropy value is independent of the ordering of the sites. Thus, higher values of entropy are intimately tied to notions of disorder and complexity and have been used productively across a wide range of disciplines.

### 5.2.3 Definition of relative-entropy

The concept of relative entropy, or Kullback-Liebler distance, is used to measure the 'distance' between two distributions of random variables. One way to think of the relative entropy $D(P\|Q)$ between two random variables P and Q is to view $D(P\|Q)$ as a measure of *inefficiency* associated with assuming that the distribution is Q, when in fact the true distribution is P [13, 66]. It is defined as

$$D(P\|Q) = \sum_{i=1}^{N} p_i \ln \frac{p_i}{q_i} \tag{5.1}$$

In our comparisons, we use the symbol Q to represent the all cancer empirical distribution, whereas P will represent a specific primary cancer type. Thus, the notion of relative entropy quantifies the relative inefficiency of using the all cancer distribution instead of the more targeted and informative primary cancer type.

## 5.3 Results

### 5.3.1 Distribution of metastatic tumors

Figure 5.1 shows the tumor distribution for all cancers collected from 3827 patients, with a total of 9484 metastatic tumors distributed over 30 distinct anatomical sites. On the left we show the histograms, normalized so that the total area under the bars is one, hence the distribution represents the probability mass function associated with all cancers. On the right we show the same data plotted on a log-log plot to more clearly bring out the fact that there is a power-law region, where the distribution follows the form $p(x) \sim x^{-\alpha}$, with $\alpha = 1.46$, obtained using maximum likelihood estimators, and a goodness-of-fit criterion for the optimal range over which the power-law holds [49]. We note that power-law distributions arise in other contexts, most relevant might be the distribution of edges from nodes on the World-Wide-Web [5]. The analogy of web-surfing from site-to-site and modeling cancer progression as a random walk process from site-to-site has been used fruitfully in [54, 55]. We caution, however, that the amount of data available from the worldwide web is orders of magnitude larger than that available from our autopsy study, large as it is. The panels shown in Figure 5.2a-l break the data of Figure 5.1 into 12 groupings associated with 12 major primary cancer types (a. Lung; b. Breast; c. Prostate; d. Colorectal; e. Pancreatic; f. Ovarian; g. Bladder; h. Cervical; i. Skin; j. Stomach; k. Uterine; l. Kidney) and the ensemble metastatic distributions associated with each. Each of the empirical distributions

shows a clear power-law range (details are described in Figure caption), each with a distinct power-law exponent and approximate range of validity.

### 5.3.2 Metastatic entropy for 12 major cancer types

Because of well known difficulties inherent with pinning down precise values for power-law exponents, we do not use their value for comparative purposes. For the purposes of quantifying the complexity associated with each primary cancer type, we calculate the 'metastatic entropy' associated with each, given by the formula $H_N = -\sum_{i=1}^{N} \sigma_i \ln(\sigma_i)$, where $\sigma_i$ represents the proportion of metastatic tumors found at anatomical site 'i', for a given primary cancer type. The constraints are given by $0 \leq \sigma_i \leq 1, (i = 1, \ldots, N), \sum_{i=1}^{N} \sigma_i = 1$. It should be intuitively clear that an increase in complexity is associated with two distinct features associated with each of the distributions: (i) the total number of sites, N, at which metastatic tumors are found, and (ii) relatively 'flat' distributions, meaning that the probabilities of spreading to each site are more equally probable than what a 'steep' distribution would show. Both of these factors play an important role in the entropy values. Table 5.1 shows the value of the metastatic entropy for each of the 12 cancer types, as well as the all cancer aggregated data. The first column lists the primary cancer type, the second column lists the number of sites, N, over which the metastatic tumors are distributed, while the third column lists the metastatic entropy associated with the empirical distributions shown in Figures 5.1 and 5.2. We list the sites according to the descending values of the entropy shown in column 3, thus Skin (2.9945), Breast (2.7798), Kidney (2.7554), and Lung (2.7453) all have entropy values

higher than the value for all cancers combined (2.7136), which we use as a benchmark for comparisons. The cancer type with the lowest entropy value is Prostate (2.0960), consistent with the relatively small number of sites to which it distributes ($N = 21$), and the relatively sharp drop in the empirical distribution shown in Figure 5.2c. It is useful to compare this distribution with Skin, shown in Figure 5.2i, which has more sites to which it distributes itself ($N = 30$), and has a distinctly flatter distribution to those sites. For ovarian cancer, whose entropy is relatively low (2.3275), we have grouped large intestine, small intestine, diaphragm, ovary, omentum, and peritoneum all as one site which we call 'peritoneal cavity', due to the fact that metastases in each of these regions likely represent random spread of disease within an anatomically connected region, as opposed to hematogenously seeded metastases.

### 5.3.3 Relative-entropy between each primary cancer type and the aggregate entropy associated with all cancers

Columns 4 and 5 in Table 5.1 show the Kullback-Liebler divergence between each cancer type and the all cancer category, as detailed in section 5.2.3. We use '$Q$' as the all cancer distribution, while '$P$' is the distribution associated with each specific cancer type. While the value of entropy shown in column 3 is independent of the ordering in which the sites are listed, the K-L divergence is not. In column 5 we calculate this quantity using the P distribution and the Q distribution arranged in decreasing order in each case, as shown in Figure 5.3. This way of comparing the distributions focuses on the 'shape' of the distribution, i.e. the

Table 5.1: **Entropy table for each cancer type and for all cancers grouped together**

| Primary Cancer Type | N | Entropy | K-L Divergence | K-L Divergence (Site Specific) |
|---|---|---|---|---|
| Skin | 30 | 2.9945 | 0.0758 | 0.1373 |
| Breast | 27 | 2.7798 | 0.0329 | 0.0759 |
| Kidney | 27 | 2.7554 | 0.0549 | 0.1352 |
| Lung | 27 | 2.7453 | 0.0360 | 0.1097 |
| All | 30 | 2.7136 | 0.0000 | 0.0000 |
| Stomach | 28 | 2.6099 | 0.0213 | 0.1191 |
| Uterine | 24 | 2.5709 | 0.0339 | 0.1459 |
| Pancreatic | 26 | 2.5540 | 0.0375 | 0.1392 |
| Colorectal | 28 | 2.4686 | 0.0351 | 0.0821 |
| Cervical | 26 | 2.3696 | 0.0546 | 0.0979 |
| Ovarian | 21 | 2.3275 | 0.0684 | 0.3416 |
| Bladder | 22 | 2.2301 | 0.0957 | 0.1477 |
| Prostate | 21 | 2.0960 | 0.1620 | 0.2750 |

Entropy table for each cancer type and for all cancers grouped together as one. First column lists the number of metastatic sites for that cancer type; second column lists the computed entropy value; third column lists the Kullback-Liebler divergence between that cancer type and the all cancer group, as compared in descending order for each; fourth column lists the K-L divergence between that cancer type and the all cancer group compared on a site specific basis. See text for more details.

rate at which it drops to zero, rather than the actual sites to which the disease spreads. As Table 5.1 column 4 indicates, the K-L divergence between Prostate and 'All' is the highest (0.1620), indicating that its shape is most different from the all cancer category. By contrast, Stomach cancer has the smallest K-L divergence from the all cancer group (0.0213), making it in this sense, the most similar to the aggregate. Column 5 in Table 5.1 shows the K-L computations between each of the cancer types and 'All' on a site specific basis, as shown in Figure 5.4. Here, we list the sites in decreasing order according to the all cancer category, meaning that the comparative histogram heights for each of the specific primary cancers

generally are not arranged in strictly decreasing order. Thus, on this site-specific way of computing the K-L divergence, Ovarian cancer (0.3416) and Prostate cancer (0.2750) have the largest values, making them the most distinct from the all cancer aggregate on a site-by-site comparison. By contrast, Breast cancer (0.0759) and Colorectal cancer (0.0821) have the smallest values of site specific K-L divergence, meaning these are the most similar to the all cancer aggregate.

## 5.4 Discussion

The approach taken in this chapter is different than in the previous chapters. The analysis was done not directly on the Markov model, but instead on the distribution that the model was built from. The entropy values that were computed allowed us to rank order the cancer types in terms of their predictability and also in terms of their similarity to general cancer. In addition to this, we also calculated the relative entropy for the distributions by matching up both in decreasing order and not by their site listings. This allowed us to analyze how similar the shape of the graphs were as opposed to the values contained within them.

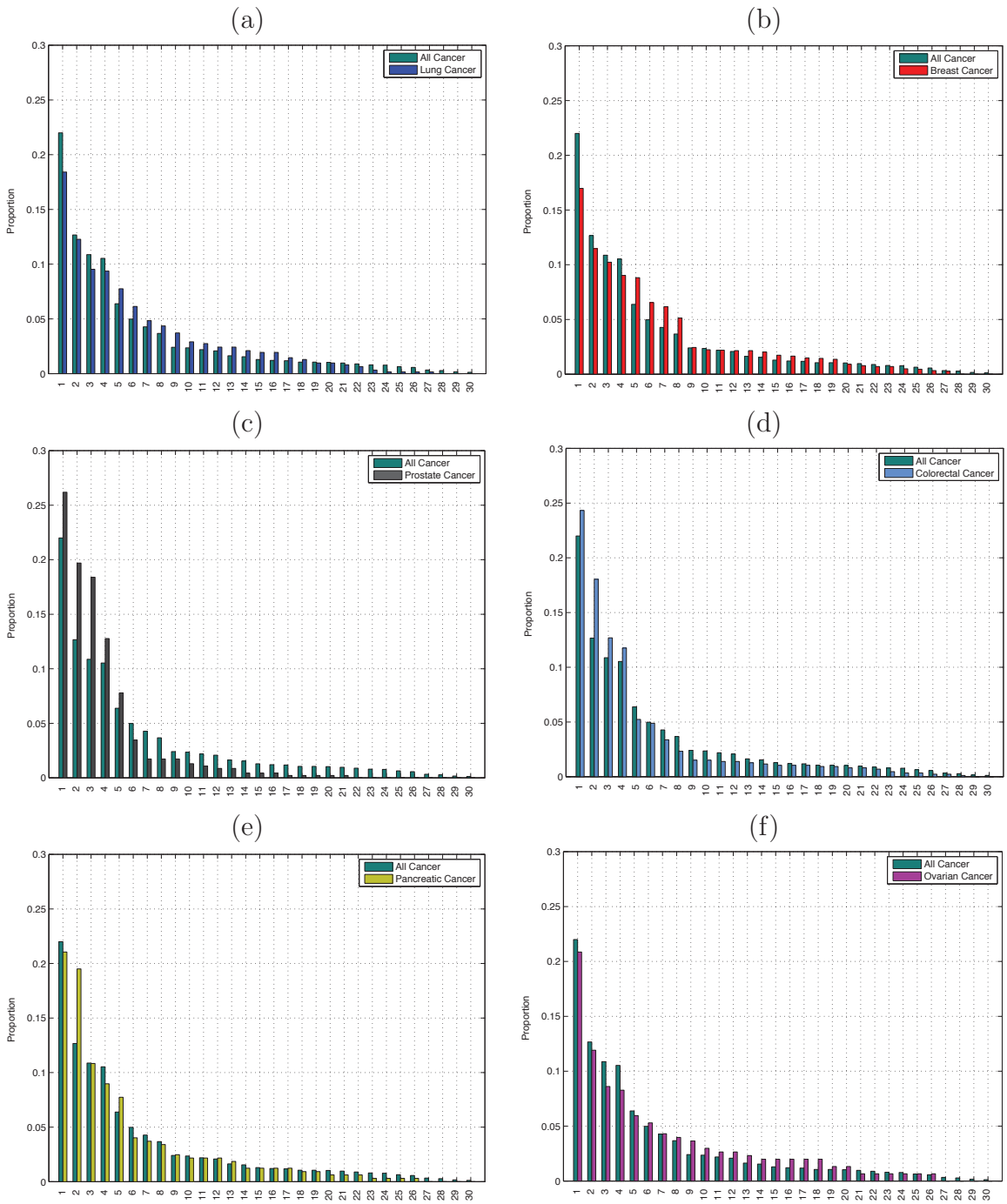


Figure 5.3: Distributions compared with all cancer (non-site specific)
Histograms of distribution of metastatic tumors for primary cancers compared with distribution of all cancer. Data is plotted in descending order for each distribution, hence is not site specific. (a) Lung cancer; (b) Breast cancer; (c) Prostate cancer; (d) Colorectal cancer; (e) Pancreatic cancer; and (f) Ovarian cancer. (*Continued on the following page.*)

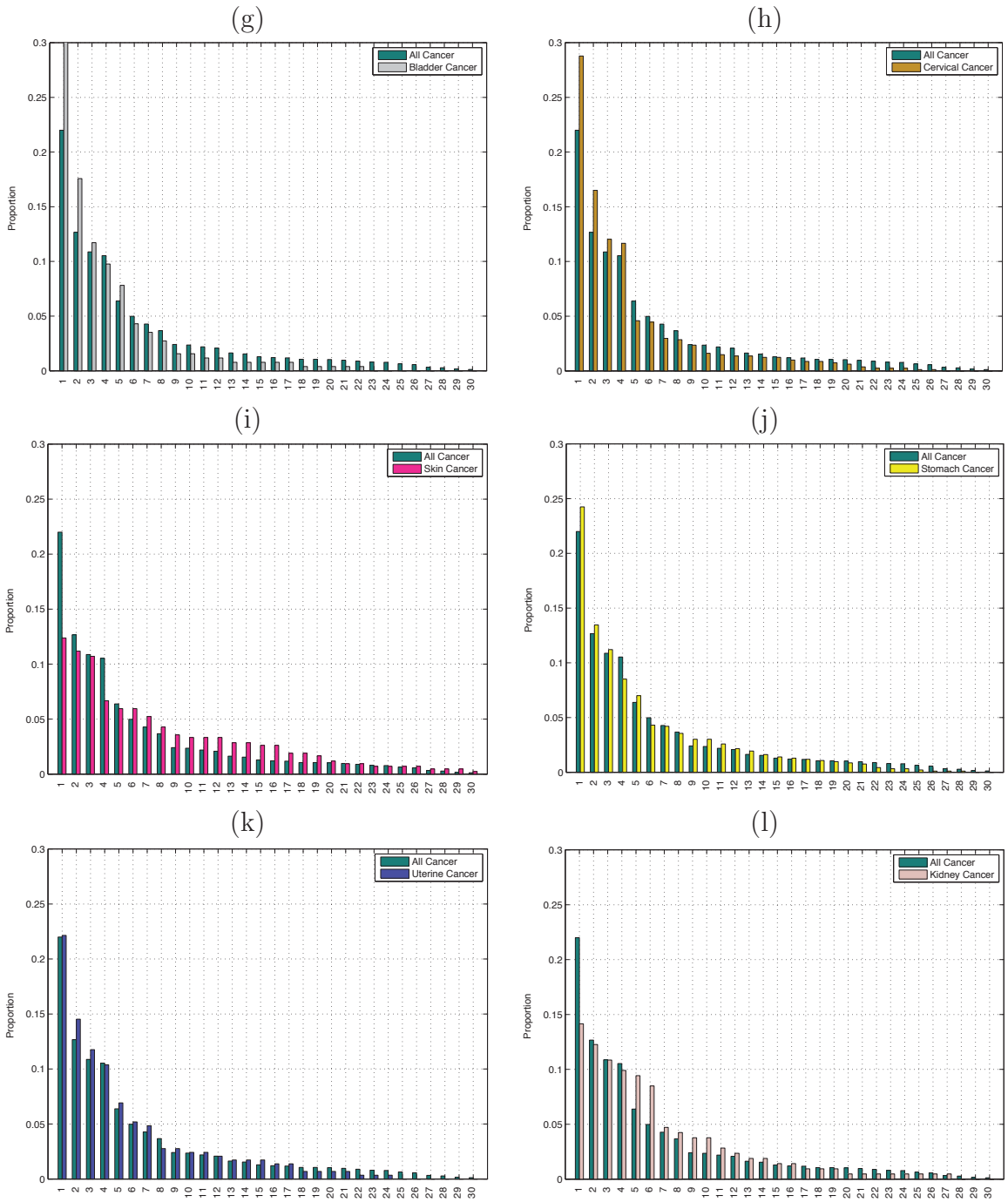


Figure 5.3 (*continued.*)

Histograms of distribution of metastatic tumors for primary cancers compared with distribution of all cancer. Data is plotted in descending order for each distribution, hence is not site specific. (g) Bladder cancer; (h) Cervical cancer; (i) Skin cancer; (j) Stomach cancer; (k) Uterine cancer; and (l) Kidney cancer.

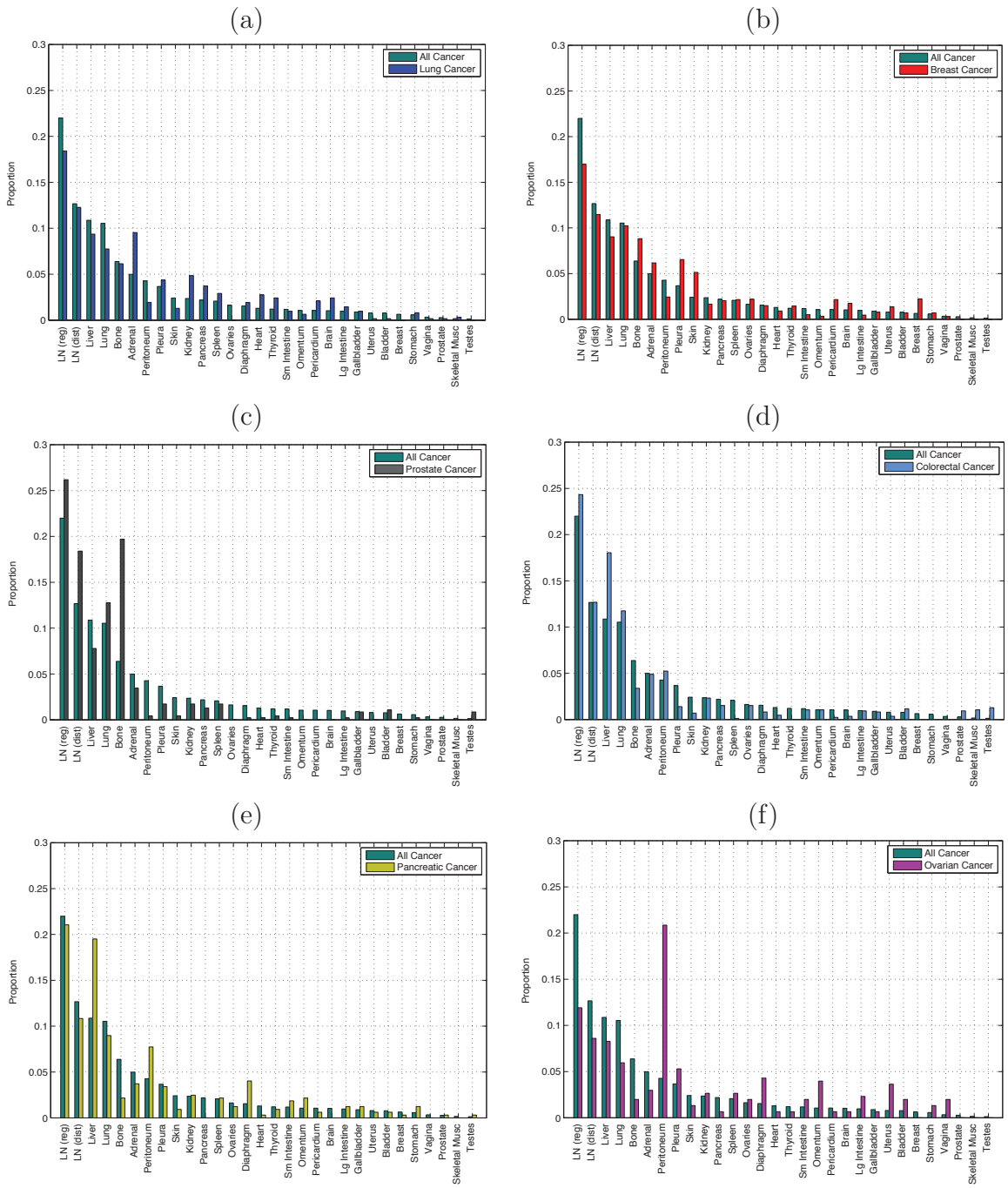


Figure 5.4: Distributions compared with all cancer (site specific)
Site specific histograms of distribution of metastatic tumors for primary cancers compared with distribution of all cancer. Data is plotted according to sites in descending order corresponding to the all cancer distribution. (a) Lung cancer; (b) Breast cancer; (c) Prostate cancer; (d) Colorectal cancer; (e) Pancreatic cancer; and (f) Ovarian cancer. (*Continued on the following page.*)

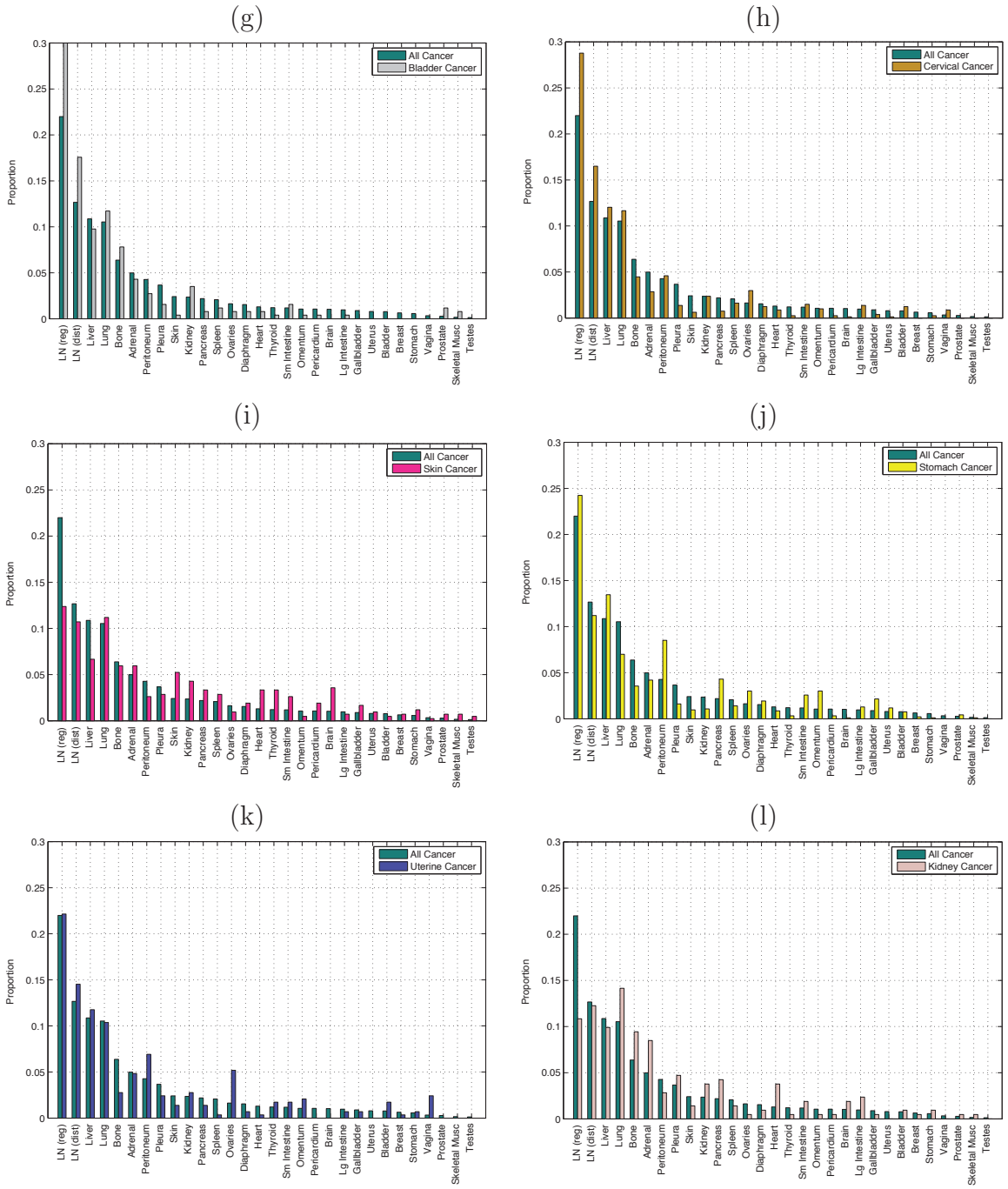


Figure 5.4 (*continued.*)

Site specific histograms of distribution of metastatic tumors for primary cancers compared with distribution of all cancer compared on a site-specific basis. Data is plotted according to sites in descending order corresponding to the all cancer distribution. (g) Bladder cancer; (h) Cervical cancer; (i) Skin cancer; (j) Stomach cancer; (k) Uterine cancer; and (l) Kidney cancer.

# Chapter 6

# Discussion

The computational model we develop and discuss in this Ph.D. thesis is an ensemble based Markov chain/random walk model of disease progression in which we use a stochastic transition matrix with entries that are (approximately) normally distributed. The model can help us quantify pathways of progression for lung cancer, and can be used as a baseline model in which to compare more targeted models which use correlations among sites making up the ensemble (i.e. the individual patients making up the ensemble), and use timescale information on disease progression. The model underscores the importance of the complex and heterogeneous nature of the connections among the many potential metastatic locations and bolsters the case for a fairly complex view of the importance of a whole host of subtle connections among sites that may or may not produce clinically detectable tumors, but that seem crucial in the eventual detailed understanding of cancer progression. We believe this autopsy based ensemble study gives important baseline quantitative insight into the structure of cancer progression networks that will be useful for future comparisons. Three key findings based on the model are:

(i) Metastatic sites can be classified into 'first-order' and 'second-order' sites based on the comparative values of the one-step vs. two-step transition probabilities. This allows us to lay out the layers of progression from lung to a given site, such as liver, shown in Figure 2.9 which lays the groundwork for a complete probabilistic classification of all pathways from primary tumor sites to metastatic locations;

(ii) The classification and quantification of 'self-seeding' transition values gives us a network based interpretation of some recent biological insights [39] that will be the focus of a separate study on probabilistic mechanisms of multidirectionality;

(iii) Model based mean first-passage times give us relative time information (based on average passage time to regional lymph nodes) about progression that can be used for future comparisons with datasets that contain time progression histories.

Our model depicts cancer progression as effectively a multistep (two-step), multidirectional, stochastic process, spreading probabilistically from site to site in individual patients, but filling out a well-defined and predictable metastatic tumor distribution for large ensembles of patients. This stable, robust, and predictable ensemble tumor distribution available over large autopsy datasets is exploited to build a Markov transition matrix for lung cancer progression. We identify the top unidirectional and multidirectional metastatic pathways of primary lung cancer by means of a probabilistic comparison of all two-step paths emanating from the lung. The results support the view that multidirectional pathways play an important role in cancer progression. We identify 3 main mechanisms of multidirectionality needed to obtain consistency with ensemble autopsy data: (i) primary tumor self-seeding,

(ii) reseeding of the primary tumor from a metastatic tumor, and (iii) metastasis reseeding. Of these, the most important are metastasis reseeding of the lymph nodes (both regional and distant) and adrenal gland and primary lung reseeding via the regional lymph nodes. Also significant is metastasis reseeding of the liver and primary self-seeding of the lung, but neither seem to be as significant as passage of the disease through the regional lymph nodes.

While very few patients die from their first metastasis, the characterization of the first metastatic site as a spreader or sponge yields important insights into metastatic pathway selection and the determination of progression timescales for patients. The model may have implications for decisions surrounding surgical resection of oligometastatic disease [76] as one might predict different outcomes for patients whose solitary site of disease is a sponge or spreader. Historically, resection of isolated adrenal metastasis has entered clinical practice in lung cancer, and removal of this spreader site has benefited patients [7]. Conversely, there has never been an established role for resection of isolated liver metastasis, a sponge site, despite there being a track record of success doing this in colon cancer [1, 24, 26, 35, 64].

A careful inspection of the top two-step pathways supports the dominance of unidirectional metastatic spread over multidirectional processes, which perhaps explains why the prevailing historical view is one of unidirectional spread [78]. However, we should emphasize that our search algorithm for a Markov transition matrix could not converge to any solution when we constrained it so that multidirectional edges were ruled out but did converge consistently to an ensemble of transition matrices when unconstrained so that all possible paths were allowed. In other words, we were not able to find a Markov transition matrix

that produced a steady-state consistent with the autopsy data unless multidirectional edge connections were allowed. Therefore, we stress the viewpoint that multidirectional processes play a key role in pathway selection and timescale determination for metastatic lung cancer.

Quantification of cancer entropy provides a framework for understanding the relative metastatic risks particular to a given tumor type. We anticipate that it will be possible to expand the application of these data to make clinical predictions and guide cancer care. An understanding of cancer entropy has important clinical implications for establishing the standards for clinical diagnosis of cancer, in guiding predictions from the growing field of radiomics, in selecting appropriate use of local therapies for oligometastatic disease, and in guiding drug development. We will explore each of these potential uses of the data below.

While the model described above is quite consistent with clinical experience, there are a number of limitations to the way in which cancer entropy is presented in this model. The first is that the model applies to a cancer primary in ensemble populations as opposed to individuals [19]. Within patients, one may have low entropy tumors, with lung as the sole metastatic site, while another may have 7 or 8 different metastatic sites. Thus the best uses of the model will be in creating generalized approaches to specific diseases. Future derivatives of this model may however be applicable to individuals whose burden of metastatic disease may be thought of as having low or high entropy with alternate approaches for each situation.

The second limitation of the model is that it groups metastasis according to organ of involvement. A patient with metastasis to the skull, pelvis, rib and femur is described as having 1 site of metastasis (bone) and therefore is scored lower for entropy that a patient

with 2 metastasis, one to the liver and the other to a portal lymph node. The information presented is thus ideally suited for understanding the implications of organ specific interventions bone directed radiopharmaceuticals, hepatic artery directed chemotherapy, etc. Any future model that uses these data to inform other forms of local therapies such as stereotactic radiation will need to consider information about the total number of metastatic sites.

Despite this work focusing mainly on the metastatic progression of a primary cancer type, the usefulness of it does not stop there. Future efforts can be made to tie this model together with multiple other models to present a more complete view of cancer in the body as opposed to one aspect. For example, one model can simulate the growth of a tumor, the next can simulate the mutation of the cancer cells, our model can predict the metastatic progression, and a final one can simulate the formation of individual metastatic sites. More models can be incorporated or omitted to model cancer on a smaller, more intricate scale, such as protein markers, signaling pathways, and drug reception/retention. One model alone cannot solve the problem of cancer, but many models working together provides a much better framework for making useful and relevant predictions to help better the medical community.

# Bibliography


[1] Abdalla, E., Vauthey, J., Ellis, L., Ellis, V., Pollock, R., Broglio, K., Hess, K., and Curley, S. Recurrence and outcomes following hepatic resection, radiofrequency ablation, and combined resection/ablation for colorectal liver metastases. *Annals of Surgery 239* (2004), 818–827.

[2] Aguirre-Ghiso, J. On the theory of tumor self-seeding: implications for metastatic progression in humans. *Breast Cancer Research 12* (2010), 304.

[3] Ashworth, T. A case of cancer in which cells similar to those in the tumors were seen in the blood after death. *Australian Medical Journal 14* (1869), 146.

[4] Balthrop, J., Forrest, S., Mewmann, M., and Williamson, M. Technological networks and the spread of computer viruses. *Science 304* (2004), 527–529.

[5] Barabási, A., and Albert, R. Emergence of scaling in random networks. *Science 286* (1999), 509– 511.

[6] Bethge, A., Schumacher, U., Wree, A., and Wedemann, G. Are metastases from metastases clinically relevant? computer modeling of cancer spread in a case of hepatocellular carcinoma. *PLoS One 12:e35689* (2012).

[7] Bretcha-Boix, P., Rami-Porta, R., Mateu-Navarro, M., Hoyvela-Alonso, C., and Marco-Molina, C. Surgical treatment of lung cancer with adrenal metastases. *Lung Cancer 27* (2000), 101–105.

[8] Butler, T., and Gullino, P. Quantitative cell shedding into efferent blood of mammary adenocarcinoma. *Cancer Research 35* (1975), 512–516.

[9] Chambers, A., Groom, A., and MacDonald, I. Dissemination and growth of cancer cells in metastatic sites. *Nature Reviews Cancer 2* (2002), 563–573.

[10] Chen, L., Blumm, N., Christakis, N., Barabási, A., and Deisboeck, T. Cancer metastasis networks and the prediction of progression patterns. *British Journal of Cancer 101* (2009), 749–758.

[11] Comen, E., Norton, L., and Massagué, J. Clinical implications of cancer self-seeding. *Nature Reviews Clinical Oncology 8* (2011), 369–377.

[12] Corbin, K., Hellman, S., and Weichselbaum, R. Extracranial oligometastases: A subset of metastases curable with stereotactic radiotherapy. *Journal of Clinical Oncology 31* (2013), 1384–1390.

[13] Cover, T., and Thomas, J. *Elements of Information Theory, 2nd Ed.* Wiley-Interscience, 2006.

[14] Cristofanilli, M., Budd, T., Ellis, M., Stopeck, A., Matera, J., Miller, C., Reuben, J., Doyle, G., Allard, J., Terstappen, L., and Hayes, D. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *New England Journal of Medicine 351(8)* (2004), 781–791.

[15] Cristofanilli, M., Hayes, D., Budd, G., Ellis, M., Stopeck, A., Reuben, J., Doyle, G., Matera, J., Allard, W., Miller, M., Fritsche, H., and Hortobagyi, G. Circulating tumor cells: A novel prognostic factor for newly diagnosed metastatic breast cancer. *Journal of Clinical Oncology 23(7)* (2005), 1420–1430.

[16] Crutchfield, J., and Young, K. Inferring statistical complexity. *Physical Review Letters 63* (1989).

[17] Csiszar, I. Why least squares and maximum entropy: An axiomatic approach to inference for linear inverse problems. *Annals of Statistics 19(4)* (1991), 2032–2066.

[18] Diaconis, P. The markov chain monte carlo revolution. *Bulletin of AMS 46(2)* (2009), 175–205.

[19] DiSibio, G., and French, S. Metastatic patterns of cancers: Results from a large autopsy study. *Archives of Pathology & Laboratory Medicine 132* (2008), 931–939.

[20] Doucet, A. *Sequential Monte Carlo in Practice*. Springer-Verlag, 2001.

[21] Edelman, E., Guinney, J., Jen-Tsan, C., Phillip, G., Febbo, P., and Mukherjee, S. Modeling cancer progression via pathway dependencies. *PLoS Comp Biology 4:e28* (2008).

[22] Ewing, J. *Neoplastic Diseases: A Textbook on Tumors*. W.B. Saunders, 6th Ed., 1929.

[23] Fidler, I. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nature Reviews Cancer 3* (2003), 453–458.

[24] Fong, Y., Cohen, A., Fortner, J., Enker, W., Turnbull, A., Coit, D., Marrero, A., Prasad, M., Blumgart, L., and Brennan, M. Liver resection for colorectal metastases. *Journal of Clinical Oncology 15* (1997), 938–46.

[25] Gamerman, D., and Lopes, H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC Publishing, 2006.

[26] Garden, O., Rees, M., Poston, G., Mirza, D., Saunders, M., Lederman, J., Primrose, J., and Parks, R. Guidelines for resection of colorectal cancer liver metastases. *Gut 55 (Suppl 3)* (2006), iii1–iii8.

[27] Goh, K., Cusick, M., Valle, D., Childs, B., Vidal, M., and Barabàsi, A. The human disease network. *Proceedings of the National Academy of Science 104* (2007), 8685–8690.

[28] Golub, G., and Van Loan, C. *Matrix Computations*. Johns Hopkins U. Press, 1996.

[29] Grinstead, C., and Snell, J. *Introduction to Probability, 2nd Ed.* American Mathematical Society, 2011.

[30] Gzyl, H. Maximum entropy in the mean: A useful tool for constrained linear problems. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 22nd International Workshop* (2003), C. Williams, Ed., American Institute of Physics, p. CP659.

[31] Gzyl, H., and Velasquez, Y. Reconstruction of transition probabilities by maximum entropy in the mean. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 21st International Workshop* (2003), R. Fry, Ed., American Institute of Physics, p. CP617.

[32] Haustein, V., and Schumacher, U. A dynamic model for tumor growth and metastasis formation. *Journal of Clinical Bioinformatics 2:11* (2012).

[33] Hoover, H., and Ketcham, A. Metastasis of metastases. *The American Journal of Surgery 130* (1975), 405–411.

[34] Hsieh, H., Marrinucci, D., Bethel, K., Curry, D., Humphrey, M., Krivacic, R., Kroener, J., Kroener, L., Ladanyi, A., Lazarus, N., Kuhn, P., Bruce, R., and Nieva, J. High speed detection of circulating tumor cells. *Biosensors and Bioelectronics 2(10)* (2006), 1893–1899.

[35] Hughes, K., Simon, R., Songhorabodi, S., Adson, M., Ilstrup, D., and Former, J. Resection of the liver for colorectal carcinoma metastases: a multi-institutional study of indications for resection. *Surgery 103* (1988), 278–88.

[36] Iwata, K., Kawasaki, K., and Shigesada, N. A dynamical model for the growth and size distribution of multiple metastatic tumors. *Journal of Theoretical Biology 203* (2000), 177–186.

[37] Jaynes, E. *Information Theory and Statistical Mechanics, in Brandeis Lectures in Theoretical Physics, Vol. 3. Ed.* W.A. Benjamin Inc. New York, 1963.

[38] Kalnay, E. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge UK, 2003.

[39] Kim, M.-Y., Oskarsson, T., Acharyya, S., Nguyen, D., Xiang, H., Norton, L., and Massagué, J. Tumor self-seeding by circulating tumor cells. *Cell 139* (2009), 1315–1326.

[40] Kinchin, A. *Mathematical Foundations of Statistical Mechanics*. Dover Publications, New York, 1949.

[41] Klein, C. Parallel progression of primary tumours and metastases. *Nature Reviews Cancer 9* (2009), 302–312.

[42] Kumar, V., Gu, Y., Baju, S., Berglund, A., Eschrich, S., Schabath, M., Forster, K., Aerts, H., Dekker, A., Fenstermacher, D., Goldgof, D., Hall, L., Lambin, P., Balagurunathan, Y., Gatenby, R., and Gillies, R. Radiomics: the process and the challenges. *Magnetic Resonance Imaging 30* (2012), 1234–1248.

[43] Lamelas, I. Long-term survival of lung cancer in a province of spain. *Journal of Pulmonary & Respiratory Medicine S:5* (2011).

[44] Leung, C., Oskarsson, T., Acharyya, S., Nguyen, D., Zhang, X. H.-F., Norton, L., and Massagué, J. Tumor self-seeding by circulating cancer cells. *Cell 139* (2009), 1315–1326.

[45] Marrinucci, D., Bethel, K., Bruce, R., Curry, D., Hsieh, H., Humphrey, M., Krivacic, B., Kroener, J., Kroener, L., Ladanyi, A., Lazarus, N., Nieva, J., and Kuhn, P. Case study of the morphologic variation of circulating tumor cells. *Human Pathology 38(3)* (2007), 1468–1471.

[46] Marrinucci, D., Bethel, K., Luttgen, M., Bruce, R., Nieva, J., and Kuhn, P. Circulating tumor cells from well-differentiated lung adenocarcinoma retain cytomorphologic features of primary tumor type. *Archives of Pathology & Laboratory Medicine 133(9)* (2009), 1468–1471.

[47] Molnar, J., Thornton, B., Molnar, A., Gaal, D., Luo, L., and Bergmann-Leitner, E. Thermodynamic aspects of cancer: possible role of negative entropy in tumor growth, its relation to kinetic and genetic resistance. *Letters in Drug Design and Recovery 2* (2005), 429–438.

[48] Newman, M. The structure and function of complex networks. *SIAM Review 45* (2003), 167–256.

[49] Newman, M. Power laws, pareto distributions and zipfs law. *Contemporary Physics 46* (2005), 323–351.

[50] Newman, M. Threshold effects for two pathogens spreading on a network. *Physical Review Letters 95(10)* (2005), 108701.

[51] Newman, M. The physics of networks. *Physics Today 61(11)* (2008), 33–38.

[52] Newman, M. *Networks: An Introduction*. Oxford University Press, 2010.

[53] Newman, M., Watts, D., and Strogatz, S. Random graph models of social networks. *Proceedings of the National Academy of Science 99* (2002), 2566–2572.

[54] Newton, P., Mason, J., Nieva, J., Bethel, K., Bazhenova, L., and Kuhn, P. A stochastic markov chain model to describe lung cancer growth and metastasis. *PLoS One 7:e34637* (2012).

[55] Newton, P., Mason, J., Nieva, J., Bethel, K., Bazhenova, L., Norton, L., and Kuhn, P. Spreaders and sponges define metastasis in lung cancer: A markov chain mathematical model. *Cancer Research* (2013).

[56] Nieva, J., Wendel, M., Luttgen, M., Marrinucci, D., Bazhenova, L., and Kolatkar, A. High-definition imaging of circulating tumor cells and associated cellular events in non-small cell lung cancer patients: a longitudinal analysis. *Physical Biology 9* (2012).

[57] Nordquist, L., Simon, G., Cantor, A., Alberts, W., and Bepler, G. Improved survival in never-smokers vs current smokers with primary adenocarcinoma of the lung. *Chest 126* (2004), 347–351.

[58] Norris, J. *Markov Chains*. Cambridge University Press, 1997.

[59] Norton, L. Gompertzian model of human breast cancer growth. *Cancer Research 48* (1988), 7067–7071.

[60] Norton, L., and Massagué, J. Is cancer a disease of self-seeding? *Nature Medicine 12(8)* (2006), 875–878.

[61] Okumura, Y., Tanaka, F., Yoneda, K., Hashimoto, M., Takuwa, T., Kondo, N., and Hasegawa, S. Circulating tumor cells in pulmonary venous blood of primary lung cancer patients. *The Annals of Thoracic Surgery 87(6)* (2009), 1669–1675.

[62] Paget, S. The distribution of secondary growths in cancer of the breast. *Lancet 1* (1889), 571–573.

[63] Paterlim-Brechot, P., and Benali, N. Circulating tumor cells (ctc) detection: Clinical impact and future directions. *Cancer Letters 253* (2007), 180–204.

[64] Pawlick, T., Abdalla, E., Ellis, L., Vauthey, J., and Curley, S. Debunking dogma: Surgery for four or more colorectal liver metastases is justified. *Journal of Gastrointestinal Surgery 10* (2006), 240–248.

[65] Pincus, S. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Science 88* (1991), 2297–2301.

[66] Rajni, B., and Agrawal, R. Mutual information and cross entropy framework to determine relevant gene subset for cancer classification. *Informatica 35* (2011), 375–382.

[67] Redner, S. *A Guide to First-Passage Time Processes*. Cambridge Univ. Press, 2001.

[68] Reynolds, S. Coming home to roost: The self-seeding hypothesis of tumor growth. *NCI Cancer Bulletin 8* (2011), 3.

[69] Ritchie, W., Granjeaud, S., Puthier, D., and Gautheret, D. Entropy measures quantify global splicing disorders in cancer. *PLoS Comp Biology 4* (2008), 1–9.

[70] Salsbury, A. The significance of the circulating cancer cell. *Cancer Treatment Reviews 2(1)* (1975), 55–72.

[71] Shannon, C. A mathematical theory of communication. *Bell System Technical Journal 27* (1948), 379–423, 623–656.

[72] Smerage, J., and Hayes, D. The measurement and therapeutic implication of circulating tumor cells in breast cancer. *British Journal of Cancer 94* (2006), 8–12.

[73] Stenbygaard, L., Sorensen, J., and Olsen, J. Metastatic pattern in adenocarcinoma of the lung: an autopsy study from a cohort of 137 consecutive patients with complete resection. *Journal of Thoracic & Cardiovascular Surgery 110* (1995), 1130–1135.

[74] Strogatz, S. Exploring complex networks. *Nature 410(6825)* (2001), 268–276.

[75] Tritchler, D., Sucheston, L., Chanda, P., and Ramanathan, M. Information metrics in genetic epidemiology. *Statistical Applications In Genetics and Molecular Biology 10* (2011), 1–20.

[76] Weichselbaum, R., and Hellman, S. Oligometastases revisited. *Nature Reviews Clinical Oncology 8* (2011), 378–382.

[77] Weinberg, R. *The Biology of Cancer*. Garland Science, 2006.

[78] Weiss, L. Metastasis of cancer: a conceptual history from antiquity to the 1990's. *Cancer Metastasis Review 19* (2000), 193–204.

[79] Weiss, L., and Ward, P. Cell detachment and metastasis. *Cancer Metastasis Review 2* (1983), 111–127.

[80] Wojtkiewicz, S. Uncertainty quantification in large computational engineering models. *AIAA-2001- 1455 19* (2001), 1–11.

[81] Yokota, J. Tumor progression and metastasis. *Carcinogenesis 21* (2000), 497–503.